U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Hélia Sofia da Rocha Monteiro da Costa

Estudo comparativo de abordagens ao problema de débito de transações bancárias em contas com saldo insuficiente

Segmento de Negócios

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Setembro de 2012

Hélia Sofia da Rocha Monteiro da Costa

Estudo comparativo de abordagens ao problema de débito de transações bancárias em contas com saldo insuficiente

Segmento de Negócios

Dissertação submetida à Faculdade de Ciências da Universidade do Porto para a obtenção do grau de Mestre em Engenharia Matemática

Dissertação realizada sob a supervisão do
Professor Doutor Luís Torgo
Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Setembro de 2012

Dedicatória

Ao meu marido, Henrique, e aos meus pais, Alberto e Otília.

Agradecimentos

Para a realização deste trabalho contei com a contribuição de várias pessoas e entidades, que permitiram enriquecer este trabalho e a quem devo meu agradecimento. Ao Dr. Manuel Gonçalves agradeço todo o incentivo, orientação e disponibilidade em esclarecer as minhas dúvidas, e em seu nome agradeço ao Millennium bcp a oportunidade e a confiança. Agradeço também ao Dr. André Martins e ao Dr. Carlos Alvim pela orientação, e em seu nome agradecer às respetivas equipas que aumentaram a sua carga de trabalho para que fosse possível concluir esta tese. Pela orientação e apoio, agradeço ao Professor Doutor Luís Torgo. Agradeço também aos colegas que dirigem comigo a Associação Casa do Povo de Santa Marinha do Zêzere, que aceitaram temporariamente uma nova distribuição de tarefas. Por fim, agradeço aos meus pais e ao meu marido pelo estímulo e por terem compreendido as minhas ausências.

Hélia Sofia da Rocha Monteiro da Costa

Setembro 2012

Resumo

O sistema financeiro possui desafios nas mais diversas áreas de atuação, sendo a decisão de risco de crédito uma delas.

A era digital intensificou os pagamentos *online* e por débito direto nas contas de depósitos à ordem. Os bancos têm que assegurar uma resposta imediata aos pedidos de pagamento de transações, que podem atingir milhões de pedidos por dia. Quando uma conta tem saldo insuficiente, o banco tem de decidir se paga a transação. Trata-se de um processo de decisão de pagar ou não pagar uma transação, cuja resposta deve ser imediata, não podendo exceder um prazo de 24 horas, conforme níveis de serviço estabelecidos para o Sistema de Compensação Interbancário. Exige-se pois um processo cujas decisões sejam rápidas, consistentes e objetivas e que minimize os erros cometidos e as perdas esperadas.

Atendendo que tanto o ciclo de receitas como o ciclo de pagamentos demoram um mês para estar completos (cerca de 22 dias úteis), espera-se que, no caso de a decisão ser pagar, a conta regularize nos 30 dias seguintes. Assim, o processo de decisão deve classificar o risco de crédito a curto prazo.

Desde 2005, num banco português, para um segmento específico designado de Mass-Market, existem modelos comportamentais e modelos que reproduzem e melhoram as regras de decisão de analistas e gestores de conta, sendo as decisões críticas sujeitas a avaliação humana.

Neste trabalho são construídos vários modelos de classificação, para duas classes, aprovação e recusa, atribuída de acordo com o incumprimento registado. Os modelos são então aplicados a um segmento específico, denominado de segmento de Negócios.

O processo de construção dos modelos, nomeadamente as várias abordagens seguidas, procuram minimizar os erros obtidos e maximizar o lucro da aplicação do processo.

Palavras-chave: risco de crédito, conta depósito à ordem, pagar ou não pagar, regularização, incumprimento, negócios, classificação,

Abstract

The financial system has challenges in many different business areas. Credit risk is one of those areas.

The digital era intensified online payments in individuals Demand Deposit Accounts (DDAs). Banks have to ensure a prompt answer for those payment requests, which can be millions a day. When a DDA has not sufficient balance, the bank has to decide whether to pay the debit transaction. This process is called Pay No Pay and it should provide an immediate response. According to the Financial Net Settlement System, service level requirements must not take more than 24 hours. Therefore, the response of this process should be fast, consistent, objective and as to minimize the possible errors and losses.

Regarding that customers' income and payments cycles take one month to be completed (22 working days), in case of Pay decision, it is expected that the DDA cures within 30 days. This led to a process that should predict short-term credit risk.

Since 2005, in Mass-Market segment of a Portuguese bank, behavioral and expert models are used to replicate and improve rules used by analysts and account managers. Critical decisions are submitted to human evaluation.

In this work several classification models are built, with two possible classes, approval and refusal. Credit in arrears will indicate which class should be assigned. These models are then applied to a specific segment, called Business.

The objective of each model is to minimize error and maximize profit.

Keywords: credit-risk, demand deposit accounts, Pay No Pay, regularization, default, business, classification.

Índice

Índice de quadros

Índice de figuras

Lista das abreviaturas

TXN – Transação

CART – Classification and Regression Trees

ENI – Empresários em nome individual

CCC – Contas correntes caucionadas

APR – Aprovação

REC – Recusa

1. Introdução

1.1. Decisões

"La vie est la somme de tous vos choix."
Albert Camus (1913 - 1960)

A evolução dos tempos permitiu aperfeiçoar as técnicas e os instrumentos utilizados pelo Homem numa das suas capacidades mais importantes, a de decidir.

O ato de decidir envolve sempre o conhecimento e a interpretação da informação necessária à decisão. Os exemplos deste envolvimento podem ser encontrados na pré-história, altura em que havia decisões tomadas através da interpretação dos sinais de fumo e dos sonhos. No século V a.C., em Atenas, o método da votação foi usado pela primeira vez, agrupando numa só a decisão dos vários cidadãos participantes. Em 1602, William Shakespeare, na obra "Hamlet", descreve as dificuldades em decidir sobre existir ou não existir, "to be, or not to be?".

As decisões que tomamos definem o nosso presente e o nosso futuro. Durante um simples dia tomamos milhares de decisões. Decidimos o que vestir, o que comer, para onde ir, como ir, enfim, o que fazer com cada segundo desse mesmo dia. Para tomar estas decisões, são avaliados vários fatores. A forma como são avaliados esses fatores poderá originar decisões diferentes, evidenciando assim a importância que têm os dados utilizados na decisão e a interpretação que é efetuada dos mesmos. Por exemplo, suponha-se que se decidiu sair de casa sem levar guarda-chuva porque o céu de manhã estava praticamente limpo. Entretanto à tarde, antes de regressar a casa, choveu e por isso a decisão de não levar guarda-chuva foi errada. Se tivesse sido prestada mais atenção às poucas nuvens presentes no horizonte e utilizada informação adicional, como a previsão meteorológica de chuva, a decisão seria provavelmente diferente. As consequências associadas a cada decisão devem também ser ponderadas no ato de decidir. No exemplo anterior, não ter chovido e ter levado guarda-chuva foi um pequeno transtorno. Porém, não ter levado guarda-chuva e ter chovido pode implicar problemas de saúde e/ou danos materiais.

Decidir corretamente, com rapidez e consistentemente tem motivado o desenvolvimento de métodos que tornem o processo de decisão objetivo e eficiente,

podendo mesmo automatizá-lo. Estes métodos procuram identificar padrões e com base nestes induzir qual a decisão correta.

1.2. Decisões de crédito

A decisão de crédito é uma área muito abrangente, à qual estão associados fatores humanos, sociais, económicos e financeiros. Um dos temas incluídos nesta área é decidir pagar transações (p.e. cobranças e cheques) em contas com saldo insuficiente para a concretização das mesmas. Ao longo deste trabalho designa-se este processo de decisão por Pay No Pay.

Devido à simplicidade de implementação, a decisão manual é a mais aplicada, sendo o gestor da conta ou cliente onde está a ser registada a transação o responsável por essa decisão. Contudo, trata-se de um trabalho moroso, dispendioso em recursos e excessivamente dependente de fatores humanos. Um processo totalmente automatizado, com decisões de aprovação ou recusa para todas as transações, permitiria ultrapassar estas desvantagens. Saliente-se contudo que tal automatização deve acautelar o risco de crédito, nomeadamente a criação de descobertos, o impacto na relação do cliente com o banco e os possíveis proveitos por juros e comissões. Assim, a tarefa de otimizar a automatização é um grande desafio, pois implica minimizar os erros e ao mesmo tempo maximizar os proveitos.

1.3. Segmento de Negócios

Com o objetivo de prestar um melhor serviço, as instituições bancárias segmentam os clientes, em função das suas características essenciais. O segmento alvo deste trabalho é o segmento de Negócios, que é formado por pequenas e médias empresas e empresários em nome individual (ENI), em função da respetiva dimensão. Algumas das características que distinguem este segmento dos restantes são o elevado número de transações, quer a débito, quer a crédito, o abrangente intervalo de montantes que cada transação pode assumir e o impacto negativo na relação comercial que a recusa de uma transação pode ter.

1.4. Objetivo

O desafio deste projeto é encontrar um processo automático de decisão Pay No Pay que se ajuste às características do segmento de Negócios e que ao mesmo tempo

procure sensibilizar os gestores de conta ou cliente para a cultura do risco de crédito, maximizando o lucro.

1.5. Estrutura da tese

Neste trabalho são abordados os vários passos para a construção de um processo automático de decisão Pay No Pay. No Capítulo 2 é descrito o domínio da aplicação. Aborda-se nesse capítulo o processo Pay No Pay, o segmento de Negócios, a informação disponível e apresentam-se os principais trabalhos existentes que estejam mais fortemente relacionados com a abordagem seguida nesta tese. No Capítulo 3 é apresentada uma formalização do problema de decisão Pay No Pay como uma tarefa de previsão, são descritos os métodos utilizados, assim como a metodologia experimental usada para os avaliar e comparar. No Capítulo 4 são descritas as várias abordagens ao problema e os respetivos resultados. Por fim, no Capítulo 5 são apresentadas as conclusões.

2. Descrição do domínio da aplicação

2.1. Processo Pay No Pay

Os progressos tecnológicos têm permitido a evolução de várias áreas e proporcionado novas soluções para problemas antigos. Os bancos são exemplos de instituições que têm procurado acompanhar estes avanços, aplicando-os às várias áreas de intervenção. Uma dessas áreas é a subscrição de risco, cujo objetivo é decidir crédito a curto, médio e longo prazo. A autorização para debitar transações em contas com saldo insuficiente é um exemplo de crédito a curto prazo. Designa-se por Pay No Pay o processo que consiste na decisão de pagamento ou não de transações diárias em contas com saldo insuficiente para as mesmas.

Para melhor se compreender o enquadramento do processo, na Figura 1 é apresentado um esquema que exemplifica o percurso de uma transação.

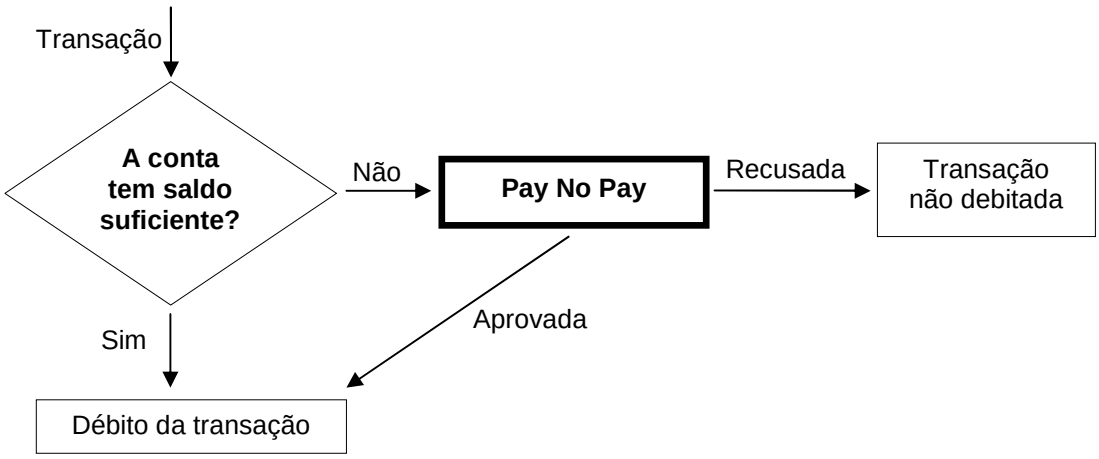


Figura 1 – Percurso de uma transação

Este processo, assumido com um serviço que é prestado ao cliente, favorece o cliente e o banco. O cliente é beneficiado, pois é-lhe autorizado o débito da transação na conta, mesmo quando a mesma não possui saldo suficiente, permitindo assim a continuidade do negócio. Nesta situação, a conta fica com saldo negativo ou agrava o saldo negativo já existente, que se designa por descoberto. O banco é beneficiado, pois cobra comissões pelo serviço. Saliente-se que a aprovação do débito é benéfica para o banco sempre que o cliente consiga posteriormente, num curto período, realizar depósitos que regularizem o descoberto. Este processo de passagem de saldos negativos a saldos positivos chama-se regularização. A prestação deste serviço não é

lucrativa para o banco quando o cliente não regulariza a conta. Nesta situação, o primeiro impacto é registado na conta à ordem, que permanece com o saldo negativo, ou ainda mais negativo, impossibilitando ao banco a cobrança dos serviços entretanto prestados. Quando o cliente não cumpre as suas obrigações para com o banco, diz-se que o cliente está em incumprimento e, com o passar do tempo, o banco deve assumir que vai perder parte ou a totalidade do crédito concedido ao cliente, não podendo utilizar esse valor para outros fins.

Para este trabalho, consideram-se vários tipos de transações, cuja decisão de aprovação do débito deve ser tomada imediatamente ou em menos de 24 horas.

2.2. Segmento alvo

As decisões são tão mais eficazes, quanto melhor se caracterizar o alvo dessas decisões.

Neste trabalho, as previsões e decisões são aplicadas ao segmento de Negócios. Este segmento é formado por pequenas e médias empresas e empresários em nome individual (ENI) e evidencia caraterísticas específicas ao nível do tipo de transações, da quantidade e frequência das transações (quer a débito, quer a crédito), dos serviços contratados, da exposição a situações de risco e do impacto para o banco em caso de abandono do cliente ou não cumprimento dos contratos.

Salientam-se como características:

- elevado número de transações diárias, quer a débito quer a crédito;
- montante das transações variável e com valores extremos bastante significativos, cuja recusa pode comprometer a atividade da empresa e cuja a errada aprovação por parte do banco pode conduzir a prejuízos;
- as transações mais frequentes são os cheques.

2.3. Transações e comissões

Neste trabalho, as transações são agrupadas em quatro grandes grupos. Por questões de confidencialidade, os grupos de transações são designados por *Ti*, com *i = 1 .. 4*. Esta segregação é justificada pelos montantes médios associados a cada tipo de transação e pelas comissões que lhes estão associadas.

De seguida são apresentados dois gráficos. Na Figura 2, o gráfico representa a proporção de cada tipo de transação no banco. Na Figura 3, para manter a confidencialidade, o gráfico com os montantes médios associados a cada grupo de transação, foi construído dividindo o montante médio de cada grupo de transações, pelo montante médio correspondente ao grupo que regista o maior valor, ou seja, dividindo pelo montante médio das transações do grupo T4.

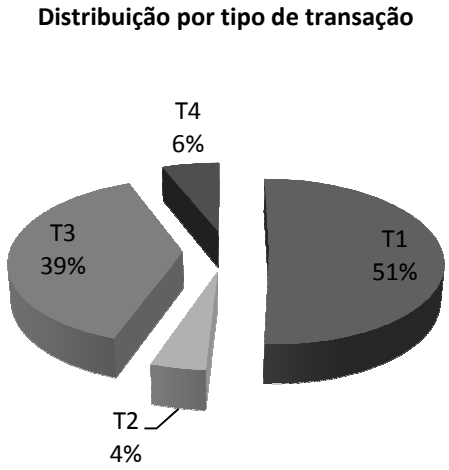


Figura 2 – Gráfico distribuição por tipo de transação

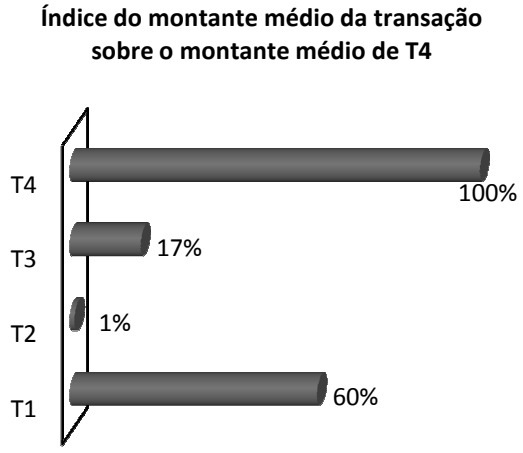


Figura 3 – Gráfico de índices de montantes médios

Além da quantidade de transações e dos montantes das mesmas, existem as comissões associadas a cada transação, que são importantes para o processo Pay No Pay. Como anteriormente referido, o banco pode cobrar diferentes comissões pela aplicação deste processo. Existem comissões de intervenção, que estão associadas à aprovação de transações. Comissões de devolução, aplicadas apenas à recusa de transações do tipo T1. As comissões por uso indevido são aplicadas a transações do tipo T2 quando a transação é debitada em uma conta que não tem saldo suficiente para a mesma. Por fim, existem as comissões ou juros de descoberto, que são aplicadas mensalmente, nas contas que registam descobertos. Por confidencialidade não são divulgados os montantes das comissões aplicadas, nem os montantes médios das transações. Porém, saliente-se que o montante médio das transações é mais de 100 vezes superior ao montante médio das comissões.

No Quadro 1 é apresentado um resumo das comissões aplicadas por tipo de transação.

	Tipo de transação			
Tipo de comissão	**T1**	**T2**	**T3**	**T4**
Intervenção	No caso de aprovação	Não	No caso de aprovação	Não
Uso indevido	Não	No caso de aprovação	Não	Não
Devolução	No caso de recusa	Não	Não	Não
Descoberto	Se a conta ficar a descoberto, pelo menos um dia.			

Quadro 1 – Transações e comissões

2.4. Informação

A decisão a tomar pelo automatismo terá que ser baseada na informação disponível até ao momento e clientes em igualdade de circunstâncias deveriam ter a mesma decisão.

Assim sendo, obter a informação que melhor perfila os clientes revela-se de extrema importância.

2.4.1. Recolha de informação

Para o desenvolvimento deste trabalho foi recolhida informação de diversos períodos temporais e de diversas fontes, internas e externas. A informação interna é referente a processos executados no banco, como o processo Pay No Pay, a concessão, cobrança e recuperação de crédito e a gestão do risco. Atendendo à diversidade e características de cada um dos processos, é possível recolher informação muito variada que permite tirar conclusões sobre o envolvimento do cliente com o banco e consequentemente da relação comercial e da rentabilidade para o banco, dos incumprimentos registados no passado, permitindo desta forma ponderar entre os lucros e ou prejuízos que cada cliente poderá originar. A componente externa, está associada a incumprimento e outras incidências de risco que são reportadas por serviços nacionais de informação. Esta informação complementa a anterior, assegurando uma completa caraterização do cliente.

2.4.2. Seleção de informação

Em primeiro lugar, identificaram-se todas as transações registadas no processo Pay No Pay durante um mês em 2011, no segmento de Negócios (83.165 registos). Para

estas transações foram identificadas as contas e os clientes associados. Recorrendo a informação armazenada em diversas bases de dados, foi recolhida informação adicional, no máximo até 12 meses antes do registo da transação. A informação disponível consistia em 115 características, das quais 6 são apenas de apoio e as restantes são relativas aos clientes, às contas e às transações a decidir. Toda a informação foi previamente transformada para efeitos de análise, incluindo os valores. No anexo A é disponibilizado um quadro com as 109 características usadas.

Todas as transações têm que ser decididas, ou seja, todas as 83.165 transações têm que ter uma decisão de aprovação ou de recusa e esta decisão só pode ser avaliada à posteriori. Para este trabalho, considera-se que a decisão de aprovação foi correta e classifica-se a transação como aprovação, se o cliente regularizou a conta nos 22 dias úteis seguintes à decisão e foi errada caso se verifique o contrário, situação em que a transação é classificada como recusa. Saliente-se no entanto que, se *M* for o montante da transação, este método de avaliação depende do saldo inicial da conta (S_i), da decisão tomada e do saldo final (S_f) nos 22 dias úteis seguintes à decisão, conforme se pode observar no esquema a seguir apresentado na Figura 4.

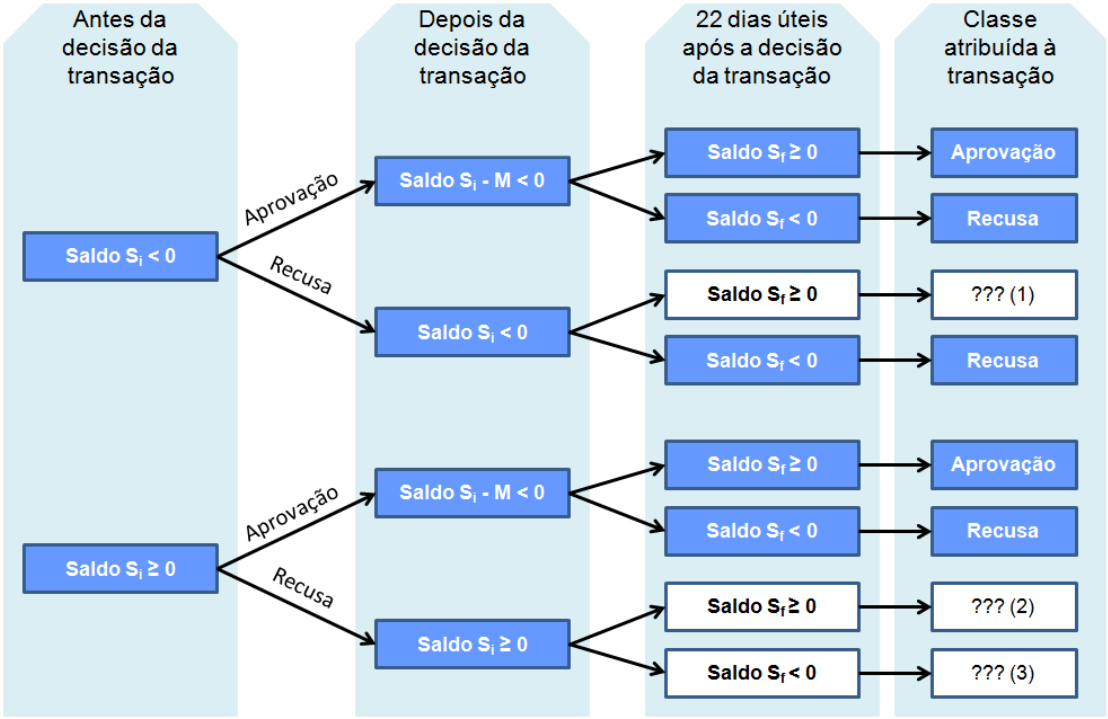


Figura 4 – Avaliação da decisão tomada

Como se pode constatar pelo esquema da Figura 4, nem sempre é possível analisar a regularização da conta quando a transação foi recusada. No caso de a transação ter sido recusada, três situações podem ser verificadas.

O primeiro caso, representado por (1) no esquema, a conta já estava a descoberto, a transação foi recusada, pelo que o descoberto não foi agravado, e o cliente regularizou a conta. Devido às frequentes alterações de saldo que ocorrem nas contas deste segmento, a análise e armazenamento dessas variações atinge uma complexidade e custos elevados. Assim, para a análise da regularização das contas, apenas são guardadas alterações de sinal no saldo e não o valor do saldo associado a cada conta. Esta restrição impossibilita qualquer conclusão quanto ao cliente regularizar a conta. Note-se que se a transação tivesse sido aprovada, o descoberto seria mais elevado e o facto de o cliente ter regularizado um montante inferior não implica que tivesse capacidade para regularizar um montante superior.

No segundo e terceiro caso, representados por (2) e (3) no esquema, a conta não estava a descoberto e, como a transação foi recusada, a conta não ficou com saldo negativo, pelo que não é possível uma análise à regularização, já que o saldo obtido nos 22 dias úteis seguintes não está relacionado com transação em causa. Neste contexto, foi decidida a exclusão destes casos do conjunto de dados usados neste trabalho.

Saliente-se que a exclusão destes casos poderia ser minimizada com o armazenamento de informação adicional, que teria como consequência o incremento de custos associados ao processo. Porém, atendendo a que o comportamento do cliente é por vezes a reação a uma ação, alguma incerteza estaria sempre associada ao processo. Por exemplo, supondo-se que foi recusada a um cliente uma transação de 20 euros, numa conta que tinha saldo zero. Durante os 22 dias seguintes, o cliente conseguiu ter no máximo um saldo de 10 euros. Numa primeira abordagem, poderíamos concluir que a decisão de recusa foi correta. Porém, se a transação fosse aprovada e fosse provocado um descoberto de 20 euros, o cliente poderia ter tomado a iniciativa de efetuar mais créditos na conta ou evitar determinados débitos nos 22 dias úteis seguintes. Nesse caso a decisão de aprovação seria a correta.

Assumindo as exclusões dos casos em que a decisão não pode ser considerada correta ou errada, o número de registos a considerar reduz-se para 68.201, conforme esquema apresentado na Figura 5.

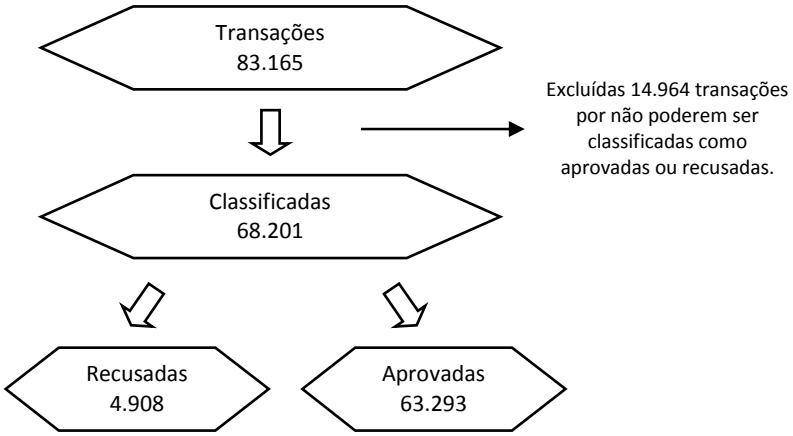


Figura 5 - Seleção de informação

Conforme se constata pelo esquema da Figura 5, 4.908 transações são recusas, ou seja 7,2% das transações, e 63.293 são aprovações, ou seja, 92,8% das transações classificadas. Torna-se agora fundamental verificar se esta proporção de casos reproduz corretamente a estrutura da população em análise.

Existem no sistema bancário transações que não requerem validação de saldo e que, por isso, são sempre debitadas nas contas. Entende-se por transações que não validam saldo aquelas que são debitadas nas contas devido a cláusulas contratuais, por obrigação legal, ou por imposição da entidade reguladora. Estas transações não são tratadas pelo processo Pay No Pay e por isso não constam dos registos utilizados neste trabalho. Porém, o facto de serem sempre debitadas nas contas permite calcular a regularização e assim ter uma aproximação para a percentagem de transações recusadas que a população deve ter. A partir de dados históricos do ano de 2011, verificou-se que a percentagem de transações que deveriam ter sido recusadas varia entre os 91% e os 93%. Com base neste resultado, não é necessária a exclusão de casos adicionais do conjunto de dados.

Como resultado do cálculo da regularização, foram adicionadas mais 4 características às 109 inicialmente existentes. Estas novas características foram recolhidas nos 3 meses seguintes à decisão e registam essencialmente a classificação atribuída a cada transação, o número de dias que a conta demorou a regularizar, para períodos de 22, 44 e 66 dias úteis, que correspondem a um, dois e três meses respetivamente.

Atendendo aos períodos que foram sendo definidos ao longo desta seção, identificando o mês das decisões como sendo o mês *M*, foram definidas três janelas temporais, a janela da observação, a janela da decisão e a janela de performance, conforme Figura 6.

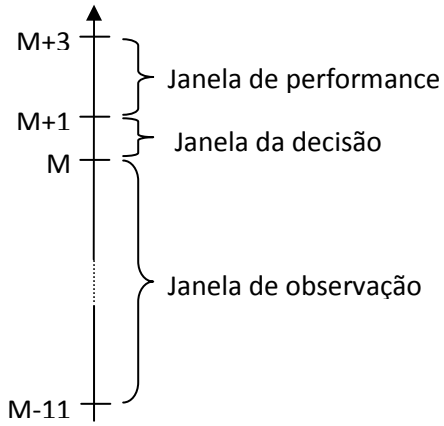


Figura 6 – Janelas temporais utilizadas

Durante o mês *M*, foram recolhidas informações sobre as transações a decidir e as contas e clientes associados às mesmas. Este período de um mês constitui a janela de decisão. Note-se que durante este período o banco manteve a sua rotina diária, pelo que as transações tiveram uma decisão manual que poderá posteriormente ser avaliada. Para os clientes e contas identificados neste período, recorreu-se à informação histórica armazenada nos doze meses anteriores, para melhor se caracterizar o cliente. Este período de doze meses constitui a janela de observação. A janela de performance é constituída pelos três meses seguintes ao dia da tomada da decisão. Neste período é analisado o comportamento do cliente, nomeadamente, se o cliente regulariza a conta, permitindo assim avaliar se as decisões tomadas foram as corretas.

2.5. Trabalhos relacionados

Vários são os estudos que tratam a decisão de autorizar o débito de transações. O que mais se aproxima do problema deste trabalho, foi publicado em 2005 e aborda também a criação de um processo automático de decisão Pay No Pay (Sousa & Costa, 2008). Esse estudo testa vários métodos, como a regressão logística, árvores de decisão e redes neuronais, para a criação de um modelo ótimo de decisão para o segmento *Mass-Market* de um banco de retalho. A construção do modelo de previsão segue duas abordagens distintas ao problema: uma usando modelos de previsão com duas classes e a outra usando modelos de previsão com três classes. As grandes diferenças entre os dois estudos estão nos segmentos alvo do processo, nos métodos usados e no tipo de abordagens efetuadas.

Existem estudos mais abrangentes, que procuram classificar o cliente com uma pontuação, que permita tomar decisões em várias áreas, como identificar quais os clientes a quem se deve propor a adesão a novos produtos, quais os clientes com tendência a abandonar o banco, quais os clientes com maior probabilidade de incumprir, ou até mesmo de cometer fraude. Trata-se de um trabalho mais exaustivo do que aquele aqui proposto e que exige a recolha de um maior conjunto de dados e a existência de um histórico mais alargado (Pliha, 2004).

Em 2005, com o objetivo de identificar se deve ser aprovado ou não um determinado crédito a um cliente, foi efetuado um estudo que procura tomar essa decisão usando *support vector machines* (Schebesh & Stecking, 2005). Neste estudo conclui-se que o método usado é facilmente aplicado a populações com diferentes proporções associadas a cada classe e com diferentes custos para cada tipo de erro cometido.

Em 2011, um estudo (Ghodselahi, 2011) aborda a utilização de um conjunto (*ensemble*) de *support vector machines* para calcular a pontuação do cliente, que tal como descrito num dos estudos referidos anteriormente, permite posteriormente tomar a decisão de aprovar ou não uma determinada operação de crédito. Os resultados obtidos confirmam a precisão deste conjunto de modelos sobre a regressão linear, árvores de decisão e *support vector machines* individuais.

3. Enquadramento teórico

3.1. Formulação do problema

3.1.1. Objetivo

Decidir transações em contas com fundos insuficientes é encontrar o equilíbrio entre o risco de cada cliente não pagar e a receita adicional por prestar esse serviço. A receita, consubstanciada em juros pelo crédito concedido e comissões pelo serviço, deve cobrir não só o custo dos fundos mas também o risco de crédito, os custos operativos e ainda remunerar os acionistas.

Pretende-se pois decidir automaticamente o maior número possível de transações em contas com fundos insuficientes, assegurando que o lucro é maximizado. Mais automatização significa decisões mais corretas e consistentes, mais rápidas e portanto mais úteis, e com maior economia. A minimização dos erros cometidos assegura não só aprovar mais transações a clientes com bom risco de crédito mas também menos descobertos a clientes de pior risco de crédito, enquanto a aplicação de comissões aumenta os lucros do banco. A decisão deverá ser efetuada transação a transação, tendo em conta a informação disponível.

Atendendo que grande parte das transações tem um ciclo de 30 dias, nomeadamente os recebimentos e os pagamentos a fornecedores e de salários, pretende-se com este trabalho prever quais as transações a aprovar, ou seja, quais as transações cujo descoberto na conta será regularizado nos 30 dias subsequentes à decisão. Na prática, pretende-se dividir o conjunto das transações em duas classes, a classe Aprovação (APR) e a classe Recusa (REC).

Demonstrar-se-á que os elementos fundamentais de decisão são por um lado o risco do cliente, isto é a probabilidade de regularizar a conta, e por outro, o modelo de *pricing* por transação e tipo de conta. Em seguida haverá que minimizar as alterações manuais da decisão dos modelos.

Para atingir este objetivo vários modelos serão testados, procurando obter aquele que minimiza o número de erros e assim o possível descontentamento do cliente e ao mesmo tempo maximiza o lucro do banco.

3.1.2. Métricas

A aplicação das diversas comissões e a criação ou o agravamento de descobertos dependem da decisão tomada. Decidir aprovar quando se deveria recusar ou, pelo contrário, recusar quando se deveria aprovar, são dois erros com consequências distintas. Aprovar erradamente origina descobertos que não serão regularizados, absorvendo parte dos proveitos do banco. Recusar erradamente, implica a perda de proveitos provenientes das comissões e pode prejudicar a relação do cliente com o banco. Com o objetivo de avaliar os vários modelos construídos serão usadas várias métricas, que procuram medir os vários tipos de erro e o lucro para cada modelo construído. Desta forma, é possível selecionar o modelo que melhor se adapta à estratégia do banco.

Para esse efeito começa-se por definir a matriz de confusão *C*, apresentada no Quadro 2.

		PREVISÃO	
		Aprovação (APR)	Recusa (REC)
REAL	Aprovação (APR)	N(APR,APR)	N(APR,REC)
	Recusa (REC)	N(REC,APR)	N(REC,REC)

Quadro 2 – Matriz de confusão C

Nesta matriz, *N(REC,APR)* representa o número de casos em que é decidida uma aprovação quando deveria ter sido decidida uma recusa. As restantes componentes da matriz são definidas de forma análoga.

Considere-se N_{Total} como o número total de decisões tomadas, pode-se então calcular a probabilidade de ser decidida uma aprovação quando deveria ter sido decidida uma recusa $P(REC, APR) = \frac{N(REC,APR)}{N_{Total}}$.

Saliente-se que:

P(APR,APR) + P(APR,REC) = P(APR,.), onde *P(APR,.)* representa a proporção de casos em que a decisão tomada deveria ser de aprovação;

P(REC,APR) + P(REC,REC) = P(REC,.), onde *P(REC,.)* representa a proporção de casos em que a decisão tomada deveria ser de recusa;

P(APR,.) + P(REC,.) = 1;

P(APR,APR) + P(REC,APR) = P(.,APR), onde *P(.,APR)* representa a proporção de casos em que a decisão tomada foi aprovação;

P(APR,REC) + P(REC,REC) = P(.,REC), onde P(.,REC) representa a proporção de casos em que a decisão tomada foi recusa;

P(.,APR) + P(.,REC) = 1.

Com base nesta matriz, reduzir o número de erros cometidos é reduzir as probabilidades *P(APR,REC)* e *P(REC,APR)*. Assim, uma das métricas utilizadas é o Erro, que é a proporção de casos na amostra em que a classe resultante da previsão não corresponde à classe real da transação, ou seja, *Erro = P(REC,APR) + P(APR,REC)*.

Note-se ainda que reduzir o erro equivale a aumentar as proporções:

$\frac{P(APR,APR)}{P(.,APR)} = P(APR \mid .,APR)$, a que chamamos a *Precision* da classe APR;

$\frac{P(REC,REC)}{P(.,REC)} = P(REC \mid .,REC)$, a que chamamos a *Precision* da classe REC;

$\frac{P(APR,APR)}{P(APR,.)} = P(APR \mid APR,.)$, a que chamamos a *Recall* da classe APR, ou especificidade;

$\frac{P(REC,REC)}{P(REC,.)} = P(REC \mid REC,.)$, a que chamamos a *Recall* da classe REC, ou sensibilidade.

Por isso, estas quatro proporções serão também usadas como métricas na avaliação do modelo.

No entanto, como referido anteriormente, cada decisão tem benefícios distintos. Por isso, define-se ainda a matriz de benefícios *B* como:

		PREVISÃO	
		Aprovação (APR)	Recusa (REC)
REAL	Aprovação (APR)	b_1	b_2
	Recusa (REC)	b_3	b_4

Quadro 3 – Matriz de benefícios B

Os coeficientes b_i dependem das comissões aplicadas e do descoberto originado pela decisão tomada. Para calcular os coeficientes b_i, considere-se x uma transação e defina-se:

- $\chi_{T1,T2,T3}(x) = \begin{cases} 1, & se\ x\ \text{é}\ do\ tipo\ T1, T2\ ou\ T3 \\ 0, & caso\ contrário \end{cases}$
- $\chi_{T1}(x) = \begin{cases} 1, & se\ x\ \text{é}\ do\ tipo\ T1 \\ 0, & caso\ contrário \end{cases}$
- $\chi_{T2}(x) = \begin{cases} 1, & se\ x\ \text{é}\ do\ tipo\ T2 \\ 0, & caso\ contrário \end{cases}$
- $\chi_{Neg}(x) = \begin{cases} 1, & se\ x\ \text{é}\ debitada\ em\ conta\ c/\ serviços\ específico\ de\ negócios \\ 0, & caso\ contrário \end{cases}$
- *ca123* – comissão de intervenção que é cobrada quando é aprovada uma transação do tipo T1, ou T2 ou T3;
- *ca1* – comissão adicional de intervenção que é cobrada quando é aprovada uma transação do tipo T1;
- *cr1* – comissão de devolução que é cobrada quando é recusada uma transação do tipo T1;
- *ca2* – comissão adicional de intervenção que é cobrada quando é aprovada uma transação do tipo T2;
- *cd* – comissão de descoberto mensal que é cobrada se a conta ficar pelo menos um dia a descoberto no mês;
- *cdn* – comissão de descoberto mensal adicional que é cobrada se a conta ficar pelo menos um dia a descoberto no mês e tiver serviços específicos para clientes de Negócios.
- *Desc(x)* – valor de descoberto provocado pelo débito da transação *x*.

Com estas variáveis, pode-se então definir:

- $b_1(x) = ca123 \cdot \chi_{T1,T2,T3}(x) + ca1 \cdot \chi_{T1}(x) + ca2 \cdot \chi_{T2}(x) + cd + cdn \cdot \chi_{Neg}(x)$
- $b_2(x) = (cr1 - ca2) \cdot \chi_{T1}(x) - ca2 \cdot \chi_{T2}(x) - ca123 \cdot \chi_{T1,T2,T3}(x) - cd - cdn \cdot \chi_{Neg}(x)$

- $b_3(x) = -0.04 \cdot \left(ca123 \cdot \chi_{T1,T2,T3}(x) + ca1 \cdot \chi_{T1}(x) - (ca123 + ca1) \cdot \chi_{T2}(x)\right) - Desc(x)$

- $b_4(x) = 0$

Assim sendo, usando a matriz de confusão C e a matriz de benefícios B podemos calcular o benefício esperado, ou seja *E[BC]*, fazendo:

E[BC] = b_1 x P(APR,APR) + b_2 x P(APR,REC) + b_3 x P(REC,APR) + b_4 x P(REC,REC).

Como b_4 é nulo, a expressão pode ser simplificada, ficando:

E[BC] = b_1 x P(APR,APR) + b_2 x P(APR,REC) + b_3 x P(REC,APR).

O benefício esperado permite identificar qual o modelo mais lucrativo para o banco.

3.1.3. Abordagens ao problema

Uma das possíveis formas de atingir o objetivo pretendido é tratar o problema como um problema de classificação. Este tipo de problemas consiste em, a partir de um conjunto de variáveis (descritores, preditores) que descrevem um caso ou objeto, encontrar uma forma sistemática de prever a que classe ele pertence. A esta forma sistemática chama-se classificador ou regra de classificação.

Por outras palavras, considere-se o conjunto *X*, que contém *n* elementos que se dispõe para a construção do classificador, isto é, *X = { x_1, x_2, …, x_n}*. Cada elemento de *X* é composto por *d* características a analisar, ou seja, *x_i = {x_{i1}, x_{i2}, …, x_{id}}*, com *i=1..n*. Seja *C* o conjunto das classes associadas a *X*, ou seja, *C = { C_1, C_2, …, C_j}*. Procura-se então uma função *D(x)*, a que se chama classificador ou regras de classificação, tal que para cada *x∈X* existe um *C_i* tal *D(x)=C_i*.

No caso concreto deste trabalho, pretende-se, para cada transação *x*, encontrar o classificador *D(x)*, tal que o resultado seja a classe Aprovação (APR) ou a classe Recusa (REC), ou seja, que possa ser usado para, dados os descritores de *x*, prever se devemos aprovar ou recusar.

Em populações em que as classes registam proporções muito diferentes, ou em que os erros cometidos têm pesos distintos como é o caso da população utilizada neste trabalho, torna-se necessário utilizar critérios mais sofisticados para passar das

previsões dos modelos para as decisões. Este conjunto de critérios, que serão discutidos nas seções seguintes, formam um modelo de decisão, que nos permitirá indicar qual a decisão a tomar (aprovação ou recusa). Como o resultado final do processo será uma decisão, este modelo é neste trabalho designado por modelo de decisão. Desta forma, a decisão final associada a cada transação resulta do encadeamento de um modelo de previsão com um modelo de decisão, conforme esquema a seguir apresentado na Figura 6.

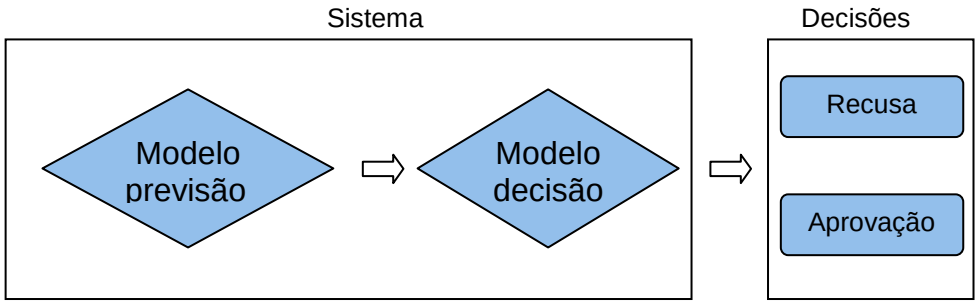


Figura 6 – Encadeamento de modelos

De seguida são descritas várias abordagens possíveis a este problema de decisão.

3.1.3.1. Minimização do erro

A primeira abordagem consiste em tentar minimizar o número de erros, ou seja, minimizar o número de casos em que:

- a decisão final obtida foi de recusa, quando deveria ser aprovação;
- a decisão final obtida foi de aprovação, quando deveria ser recusa;

Para tal, são apenas usados modelos de classificação para prever a classe da transação. Ou seja, é atribuída a classe APR ou REC com base na classe mais frequente em cada subconjunto construído pelos modelos de previsão. Nesta primeira abordagem a decisão é igual à previsão do modelo de classificação.

Os modelos de classificação que serão considerados nesta abordagem são as *Random Forests* e as árvores de decisão, que são descritos na seção 3.2. Para estes dois métodos foram testados vários parâmetros, sendo originados vários modelos.

Para avaliar e comparar os resultados, foi usada a metodologia experimental da validação cruzada, descrita na secção 3.3. Com base nos resultados obtidos, são construídas as métricas anteriormente definidas na seção 3.1.2.

Atendendo à matriz de benefícios B, definida na seção 3.1.2, espera-se que os dois tipos de erros referidos no início desta seção tenham impactos diferentes no cálculo dos custos para o banco. Por isso, a minimização do número de erros, pode não conduzir à maximização do lucro. Esta possibilidade levou à consideração da abordagem seguinte.

3.1.3.2. Maximização do lucro por definição de probabilidade limite

Atendendo que o montante médio das transações é bastante superior ao valor médio das comissões aplicadas, o erro de aprovar uma transação que deveria ser recusada é muito mais penalizador economicamente que o erro de recusar uma transação que deveria ser aprovada. Assim sendo, é expectável que o lucro do processo seja reduzido ou mesmo nulo, quando se usa como único critério a minimização do número de erros.

Nesta abordagem procura-se encontrar formas que conduzam à maximização do lucro. Para tal, usa-se um modelo de decisão que é baseado na confiança das previsões do modelo de classificação. Procura-se deste modo que as decisões de aprovação só tenham lugar quando o modelo de classificação exibe grande confiança na sua previsão. Para isso é necessário que os modelos usados sejam capazes de produzir não só uma previsão mas também a confiança na mesma, o que é o caso dos modelos de classificação considerados neste trabalho.

Concretamente, o modelo de decisão atribui a classe APR se a confiança na previsão do modelo de classificação for superior a um determinado valor de corte. Assim, poderão ser definidos diferentes modelos com base em diferentes valores de corte desta confiança. Por exemplo, se o valor de corte for 70% e a confiança numa previsão APR for de 80%, então a decisão para este caso será aprovar a transação.

Nesta abordagem foram usados os modelos de *Random Forests* e, para avaliar e comparar os resultados, foi usada a metodologia experimental da validação cruzada, descrita na secção 3.3. Com base nos resultados obtidos, são construídas as métricas anteriormente definidas na seção 3.1.2.

3.1.3.3. Maximização do lucro usando matriz de benefícios

Na sequência do raciocínio que conduziu às abordagens anteriores, ou seja, que o erro de aprovar uma transação que deveria ser recusada é muito mais penalizador economicamente que o erro de recusar uma transação que deveria ser aprovada, procura-se utilizar um novo modelo de decisão que incorpore a matriz de benefícios B, variável com cada transação, conforme definida na seção 3.1.2.

Tal como na abordagem anterior, é usado um modelo de classificação que tem como resultado a probabilidade de cada uma das classes possíveis.

O modelo de decisão leva em conta a probabilidade associada à previsão APR, *P(.,APR)*, a probabilidade associada à REC, *P(.,REC) = 1 - P(.,APR)*, e a matriz de benefícios *B*. A cada transação, será associada a classe aprovação quando aprovar é mais lucrativo que recusar. Por outras palavras, decide-se que uma determinada transação é da classe APR quando $b_1 \times P(.,APR) + b_3 \times P(.,REC) \geq b_2 \times P(.,APR) + b_4 \times P(.,REC)$. As restantes transações serão associadas à classe REC.

Nesta abordagem foram novamente usados os modelos de árvores de decisão e *Random Forests* e, para avaliar e comparar os resultados, foi usada a metodologia experimental da validação cruzada, descrita na secção 3.3. Com base nos resultados obtidos, são construídas as métricas anteriormente definidas na seção 3.1.2.

3.1.3.4. Maximização do lucro adicionando critérios *expert*

Pretende-se com esta abordagem, incluir a experiência adquirida pelos decisores manuais ou critérios de definam o risco máximo que o banco está disposto a assumir e, por esta via, melhorar os resultados da métrica Lucro definida na seção 3.1.2. A adaptação desta experiência adquirida é efetuada através da inclusão de validações simples aplicadas a uma ou mais características. Estas regras denominam-se por regras *expert*.

Tal como na abordagem anterior, é usado um modelo de classificação que tem como resultado a probabilidade de cada uma das classes.

O modelo de decisão resulta do encadeamento do modelo de decisão definido na abordagem anterior, que inclui a matriz de benefícios *B*, e as regras simples de validação de uma determinada característica.

Neste trabalho foram testadas duas regras *expert*. A primeira consiste na validação do valor a descoberto na conta caso a transação fosse aprovada. Ou seja, se o descoberto provocado for superior a um valor *S*, a decisão final da transação é REC, caso contrário é mantida a decisão obtida pela abordagem anterior. O valor *S*, por questões de confidencialidade não é divulgado neste documento. A segunda regra resulta da validação do rácio entre o descoberto provocado e o montante da transação. Se este rácio for inferior a 50%, a decisão final é aprovação, caso contrário, mantém-se a decisão obtida pela abordagem anterior. Na prática, estas regras, ajustam a decisão do modelo da Seção 3.1.3.3., para os casos em que as condições se verifiquem.

Nesta abordagem foram novamente usados os modelos de árvores de decisão e *Random Forests* e, para avaliar e comparar os resultados, foi usada a metodologia experimental da validação cruzada, descrita na secção 3.3. Com base nos resultados obtidos, são construídas as métricas anteriormente definidas na seção 3.1.2.

3.2. Métodos

Para atingir o objetivo proposto neste trabalho poderiam ser usados vários métodos, tais como as redes neuronais, regressão logística, máquinas de suporte vetorial, entre outros.

O banco atualmente utiliza árvores de decisão para o processo Pay No Pay, em alguns segmentos de clientes. A escolha desta metodologia deveu-se essencialmente aos seguintes fatores:

- ás aplicações informáticas disponíveis no banco;
- à facilidade de aplicação;
- ao facto de permitir conhecer melhor o objeto do estudo e as suas relações;
- à fácil interpretação dos resultados;
- não ser necessário conhecer as funções de distribuição associadas aos dados.

Neste contexto, foi decidido também usar as árvores de decisão (classificação) neste trabalho. Além destas, foi ainda usado o método *Random Forests*, devido às exigências do segmento de Negócios e aos resultados obtidos pelo método das *Random Forests* em outras aplicações e contextos. Como este método consiste na utilização conjunta de várias árvores de decisão, de seguida são aprofundados vários conceitos relativos a árvores de decisão e posteriormente ao método *Random Forests*.

3.2.1. Árvores de decisão

As árvores de decisão surgiram na década de 60, com Morgan & Sonquist, que construíram o primeiro algoritmo denominado de AID (Automatic Interaction Detection). Com o passar dos anos, os algoritmos foram sendo aperfeiçoados, sendo dada especial relevância ao trabalho desenvolvido por Breiman et al e o respetivo algoritmo CART (Classification And Regression Tree) (Stone, Breiman, Olshen, & Friedman, 1984) e Quinlan com o algoritmo ID3 (Iterative Dichotomiser 3) (Quinlan, 1986).

3.2.1.1. Descrição

A filosofia de funcionamento de qualquer árvore de decisão é bastante simples, visto que se baseia na estratégia de dividir para conquistar. De uma forma geral, uma árvore de decisão baseia-se na sucessiva divisão do problema em vários sub-problemas de menores dimensões, até que uma solução mais simples para cada um dos problemas possa ser encontrada. Por outras palavras, procura-se dividir sucessivamente o conjunto inicial em vários subconjuntos. Para efetuar as divisões são usadas questões cujas respostas permitem a criação de dois ou mais subconjuntos. Neste tipo de métodos, as previsões podem ser obtidas seguindo o caminho ditado pelas sucessivas questões colocadas ao longo da árvore, a que se chama nós, até que seja encontrada uma folha (subconjunto final) que conterá o resultado. Na Figura 7 é apresentado um exemplo de uma árvore de decisão, para classificação.

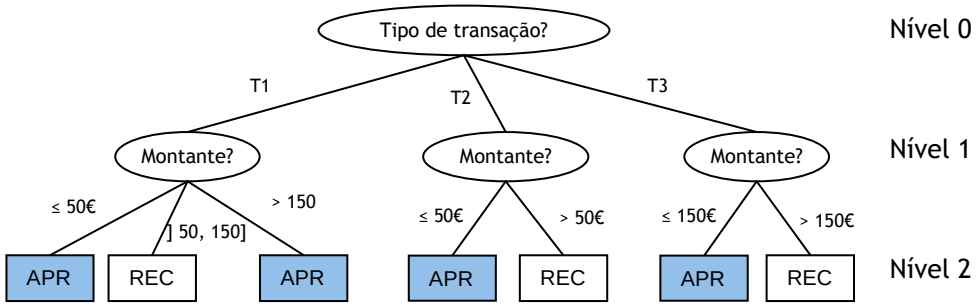


Figura 7 – Árvore de decisão

Esta árvore de decisão divide o conjunto inicial em vários subconjuntos, conforme esquema na Figura 8.

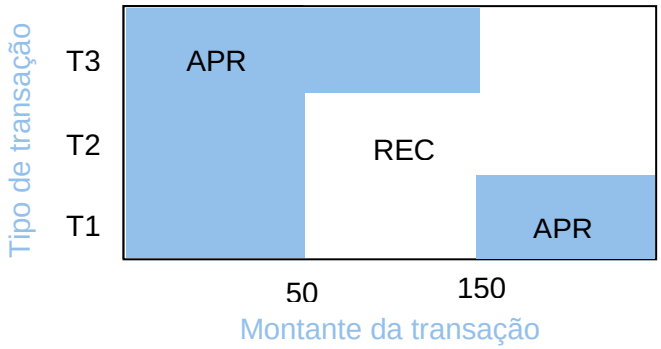


Figura 8 – Partição do conjunto inicial através de uma árvore de decisão

Tendo em conta que a criação de um caminho depende do tipo de questão ou teste que for colocado, muitas são as possibilidades para a construção de árvores de decisão. No entanto, este não é o único fator que influencia a construção de árvores de decisão.

Abaixo seguem alguns aspetos a considerar aquando da construção de uma árvore:

1. O número de subconjuntos a ser criados a partir de uma questão/teste;
2. A característica/propriedade a ser testada em cada nó;
3. Se um nó deve ser considerado um nó final ou não;
4. Que classe atribuir ao nó final.

De seguida analisar-se-á cada uma destas questões.

Quanto ao ponto 1, note-se que qualquer divisão em mais do que dois subconjuntos pode ser convertida em sucessivas divisões de apenas dois subconjuntos. Como por exemplo, a árvore da Figura 7 pode ser substituída pela seguinte árvore.

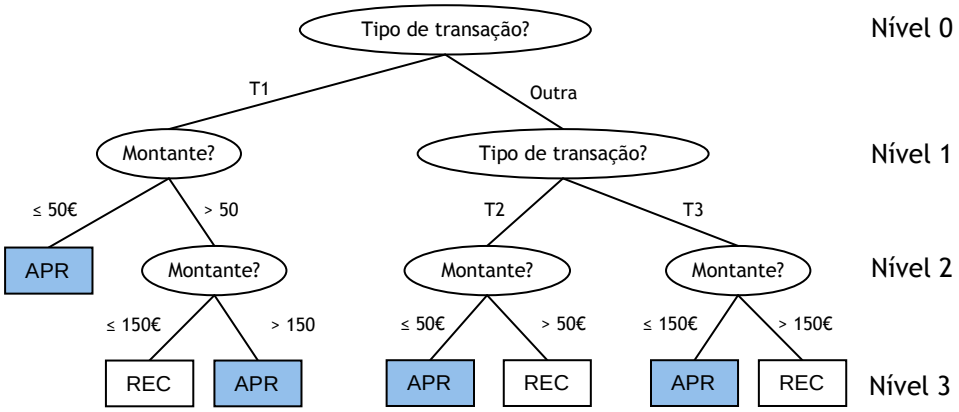


Figura 9 – Árvore de decisão binária

As árvores da Figura 9 denominam-se de “árvores binárias”. Dada simplicidade de compreensão e relativa facilidade de manuseamento deste tipo de árvores, este estudo será baseado em árvores binárias.

Os pontos 1 e 2 estão diretamente relacionados, pois o tipo de propriedade usada pode conduzir diretamente ao número de subconjuntos a ser criados. Por exemplo, uma característica que tenha como valores possíveis 0 e 1, só poderá originar dois subconjuntos diferentes, um com os elementos em que a característica assume o valor 0 e outro com os elementos em que a característica assume o valor 1. Ora, se m_i for o número de valores distintos da característica x_i, duas situações podem acontecer:

- Se a característica é nominal, as questões são do tipo “$x_i \in S$?”, onde S é um subconjunto do conjunto dos valores possíveis da característica. Nesta situação, podem ser construídas $2^{m_i-1}-1$ questões/ testes diferentes com base nesta característica.
- Se a característica for quantitativa, as questões são do tipo ”é $x_i \leq c$?”, onde c é uma constante no intervalo de valores possíveis para as características. Nesta situação, podem ser construídas no máximo m_i questões/ testes diferentes com base nesta característica ordenada.

Quanto ao ponto 2, procura-se uma característica que permita a divisão de um nó inicial em nós menos “impuros”, sendo que um conjunto será tanto mais “impuro” quantos mais elementos tiver de uma classe diferente da que foi associada ao nó. Assim sendo, torna-se fundamental definir uma medida de impureza.

Seja $P_N(C_i)$ a proporção de elementos de N que pertencem à classe C_i tal que $P_N(C_1)+P_N(C_2)+\ldots+P_N(C_j)=1$. Uma medida de impureza é uma função Φ que verifica:

1. $\Phi(P_N(C_1),P_N(C_2),\ldots,P_N(C_j))$ é máxima quando $P_N(C_1)=P_N(C_2)=\ldots=P_N(C_j)=1/j$;
2. $\Phi(1,0,\ldots,0)=\Phi(0,1,\ldots,0)=\Phi(0,0,\ldots,1)=0$;
3. Φ é uma função de simétrica $P_N(C_1),P_N(C_2),\ldots,P_N(C_j)$.

Por outras palavras, seja $i(N)$ a impureza do nó N, então $i(N)$ é máxima quando as classes estão igualmente distribuídas no nó e é mínima quando o nó contém apenas elementos de uma única classe. Algumas das medidas mais populares de impureza de um nó são:

1. Entropia: $i(N) = -\sum_j P_N(C_j) \cdot log_2\, P_N(C_j)$.

2. Gini: $i(N) = \sum_{i \neq j} P_N(C_i) \cdot P_N(C_j) = 1 - \sum_j P_N^2(C_j)$. No caso de existirem apenas duas classes esta fórmula fica $i(N) = P_N(C_1) \cdot P_N(C_2)$ e designa-se por variância.
3. Misclassificação: $i(N) = 1 - \max_j P_N(C_j)$.

Na Figura 10 pode-se comparar a variação das três medidas com a proporção dos elementos das classes no nó, para o caso de apenas duas classes.

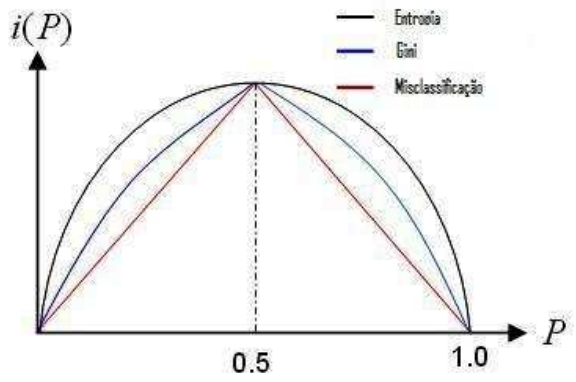


Figura 10 – Comparação entre medidas de impureza

O que se procura então é uma característica e uma questão que divida o nó inicial em nós descendentes de forma a maximizar a variação na impureza, ou seja, de forma a maximizar $\Delta i(N) = i(N) - P_E(N_E) - P_D(N_D)$, onde N_E e N_D são os nós descendentes, à esquerda e à direita, respetivamente e P_E e P_D são as proporções de elementos que vão para o nó descendente esquerdo e direito respetivamente.

Quanto ao ponto 3, pretende-se saber quando deve ser parado o crescimento de uma árvore de decisão. Este passo é importante, pois quando o processo é parado de forma incorreta, duas situações podem acontecer:

- parar demasiado cedo, provocando assim um erro elevado em cada folha, penalizando assim a performance do modelo.
- parar demasiado tarde, o que implica uma adaptação excessiva do modelo aos dados usados na sua construção (*overfitting*), que num caso extremo, pode levar à existência de um elemento por folha. Nesta situação, o resultado é excelente para os dados usados na construção do modelo, mas pode resultar em erros elevados quando aplicado o modelo a outros dados.

Para resolver esta questão existem várias soluções. Uma possível solução consiste na aplicação de métodos experimentais, como por exemplo, a validação cruzada em que experimentalmente é determinado qual o ponto de paragem que minimiza o erro.

Outra solução é a definição à priori de um valor limite, β, para a redução da impureza. Ou seja, o processo para, quando, ao adicionar mais um nó, se obtém $max_s \Delta i(s) \leq \beta$. Uma outra solução consiste em parar o processo de crescimento atendendo à complexidade do processo. Por outras palavras, para-se o processo quando for atingido o $\min(\alpha \cdot \gamma + \sum_{\text{folhas}} \text{i(N)})$, onde α é uma constante a definir e γ é o número de nós da árvore. Note-se que as duas primeiras soluções dependem apenas do último nó construído, enquanto a terceira solução depende de toda a árvore. Outra solução possível é parar o crescimento da árvore com base na análise estatística da significância associada à redução da impureza. Para analisar a significância estatística são usados testes de hipóteses como a estatística do qui-quadrado e intervalos de confiança. Ou seja, se o acrescentar de mais um nó não traz estatisticamente mais significância, então para-se o crescimento da árvore.

Uma outra forma de ultrapassar os problemas associados à paragem antecipada ou tardia do crescimento da árvore consiste na aplicação do processo de poda. Contrariamente aos processos apresentados anteriormente, o crescimento da árvore não é parado, ou seja, a árvore é construída até ao final, ou seja, até que as folhas finais tenham impureza mínima. Posteriormente procede-se à eliminação folhas ou sub-árvores cuja contribuição para a melhoria da impureza é pouco relevante de um ponto de vista estatístico.

Quanto ao ponto 4, após identificar todas as folhas da árvore, torna-se necessário decidir qual a classe a atribuir. Nos casos em que todos os elementos da folha são da mesma classe, ou seja, a impureza é zero, obviamente a classe a atribuir é essa mesma classe. Porém, da aplicação dos processos de paragem das árvores ou da poda resultam em regra folhas onde a impureza é positiva. Nestes casos, o procedimento habitual é a atribuição da classe que regista maior proporção de elementos na folha. Ou seja, escolhe-se a classe j_0 que verifica $P_N(j_0) = \max_j P_N(j)$.

3.2.1.2. Vantagens e desvantagens

A utilização de árvores de decisão tem algumas desvantagens:

- Instabilidade. Pequenas perturbações do conjunto de treino podem provocar grandes alterações no modelo aprendido.
- Fragmentação de conceitos e replicação de sub-árvores.

No entanto, as árvores de decisão são muito utilizadas devido às seguintes vantagens:

- Trata-se de um método não paramétrico e por isso não é necessário conhecer a distribuição dos dados;
- Não depende da escala das variáveis;
- Fácil interpretação dos resultados obtidos, dado que resultam do encadeamento de decisões elementares;
- É eficiente na construção de modelos;
- Robusto á presença de pontos extremos e atributos redundantes ou irrelevantes.

Estas vantagens motivaram a utilização das árvores de decisão na criação de grelhas de decisão automática no processo Pay No Pay.

3.2.2. Random Forests

Para o desenvolvimento de modelos para segmentos de clientes cada vez mais específicos e mais exigentes, leva a questionar sobre a existência de métodos que possam adequar-se melhor aos processos. Uma das soluções encontradas consiste na utilização de vários modelos em conjunto. Este tipo de solução permite resolver problemas como:

- Informação reduzida;
- Incapacidade dos algoritmos de aprendizagem para alcançar o melhor resultado, devido à complexidade computacional;
- Inadequadas representações da função preditiva devido aos pressupostos associados aos modelos.

Foram vários os modelos criados que utilizam um conjunto de árvores de decisão para previsão. A previsão deste tipo de modelos é, regra geral, a escolha do resultado mais votado atendendo às árvores criadas. Em 1996, surgiu o método de *Bagging* (Breiman, 1996a), que consistia na construção de várias árvores de decisão, em que para o crescimento de cada árvore era utilizada uma amostra por *boostrap*, sem substituição, de amostras do conjunto de treino.

Em 1998, surgiu o método da escolha aleatória dos nós (Ho, 1998). Neste método, os nós são escolhidos aleatoriamente de um conjunto mais reduzido que contém os k melhores nós. Em 1998, vários artigos são escritos sobre o espaço aleatório, que consistia na escolha aleatória de uma conjunto de características que seriam usadas na construção das árvores individuais. O método que será utilizado neste trabalho

surgiu em 2001, com Leo Breiman (Breiman, 2001) e denomina-se de *Random Forests* (Florestas Aleatórias).

As *Random Forests* são um método que consiste na construção de um conjunto de árvores de decisão, tal como descrito no ponto 3.2.1, com a única diferença de as árvores individuais não passarem pelo processo de poda (*prunning*).

A construção de *Random Forests* agrega dois conceitos:

- a escolha aleatória de amostras do conjunto de treino;
- a escolha aleatória de um subconjunto de variáveis em cada nó, durante o processo de crescimento de cada árvore individual, de onde é escolhido o teste a usar nesse nó.

A previsão final da *Random Forests* para problemas de classificação é a previsão maioritária entre o conjunto de árvores de decisão individuais construídas.

A variabilidade entre as árvores está assegurada pelas sucessivas escolhas do conjunto de treino (*bootstrap*) e pelo número de características a considerar em cada nó para selecionar a divisão a efetuar.

3.2.2.1. Descrição

Cada árvore de decisão é obtida a partir de uma fração do conjunto de registos disponíveis, escolhida aleatoriamente. A este conjunto chamamos conjunto de treino. Os restantes registos são posteriormente utilizados para testar a qualidade do modelo. A este conjunto chamamos conjunto de teste.

Assim, supondo que o número de registos disponíveis para a construção do modelo é *n* e que o número de características disponíveis é *d*, deve-se escolher para a construção das árvores individuais:

- *s* registos dos *n* disponíveis, permitindo recolocação, construindo assim o conjunto de treino. Este número é, regra geral, 2/3 de *n*.
- *m* características de entrada para determinar a decisão em cada nó da árvore. Este número *m* deve ser muito menor do que *d*.

Para cada nó da árvore, calcula-se o melhor corte a aplicar, atendendo às *m* características selecionadas. As árvores individuais são inteiramente crescidas, ou seja, não se recorre ao processo de poda.

3.2.2.2. Vantagens e desvantagens

O método das *Random Forests* tem várias vantagens, nomeadamente:

- É atualmente um dos classificadores mais exatos.
- Pode estimar a importância das variáveis para métodos de classificação.
- Gera uma estimativa interna não enviesada do erro da generalização aquando da construção do modelo.
- Aceita e trata com exatidão características que tenham dados em falta.
- Fornece um método experimental que deteta interações entre variáveis.
- Rapidez de construção, mesmo em conjuntos com um número elevado de objetos e/ou características.
- Pode ser usado para *clustering* ou identificação de *outliers*.

As principais desvantagens são:

- Dificuldade em interpretar o modelo, ao contrário das árvores de decisão.
- Tendência para *overfit* em casos de bases de dados com muito ruído.
- Enviesa os resultados quando existem características nominais com elevado número de atributos distintos. A aplicação de métodos com a permutação parcial permite resolver este problema.

3.3. Metodologia experimental

Os resultados dos vários métodos utilizados devem ser consistentes e permitir avaliar a capacidade de previsão de cada modelo. Assim sendo, a metodologia experimental utilizada reveste-se de especial importância.

Nesta seção serão descritas as metodologias experimentais utilizadas neste trabalho.

3.3.1. Validação cruzada

A validação cruzada é uma técnica utilizada na avaliação da capacidade de generalização de um modelo, a partir de um conjunto de dados.

Esta técnica é utilizada com frequência em problemas de previsão, para a análise do desempenho do modelo num conjunto de dados, ou seja, para estimar quão preciso é o modelo.

A base das técnicas de validação cruzada consiste na divisão do conjunto de dados em subconjuntos mutuamente exclusivos. Posteriormente, alguns destes subconjuntos são utilizados na estimação de parâmetros do modelo, ou na construção do modelo e designa-se por conjunto de treino. Os restantes subconjuntos, que se designam por conjunto de teste, são usados na validação do modelo.

Existem diversas formas de realizar a divisão dos dados, sendo as três mais utilizadas: o método *holdout*, o *k-fold* e o *leave-one-out*.

No âmbito deste trabalho, atendendo ao número elevado de dados, será utilizado o método *holdout*, pois é o mais indicado nestas situações.

O método *holdout* consiste em dividir de forma aleatória o conjunto total de dados em apenas dois subconjuntos mutuamente exclusivos, denominados de conjunto de treino e conjunto de teste. Estes conjuntos podem ou não ter o mesmo tamanho. Uma proporção muito comum é considerar 2/3 dos dados para o conjunto de treino e o 1/3 restante para o conjunto de teste.

No sentido de aumentar a significância estatística dos testes realizados, também é frequente repetir a divisão aleatória do conjunto de dados em subconjunto de treino e teste. Nas experiências deste trabalho esta divisão foi efetuada 10 vezes.

3.3.2. Teste de *Wilcoxon*

O teste de Wilcoxon é um teste de hipóteses, não paramétrico, usado quando se compara duas amostras dependentes, amostras emparelhadas ou medições repetidas numa mesma amostra, e se pretende analisar as diferenças entre os valores médios.

A aplicação deste teste parte dos seguintes pressupostos:

- os dados são emparelhados e provêm da mesma população;
- cada par de dados é escolhido aleatoriamente e é independente dos restantes;
- os dados são medidos numa escala de intervalos, que não precisa de ser normal.

Considere-se H_0: a diferença entre as médias é nula e H_1: caso contrário.

O algoritmo para execução deste teste é o seguinte:

1. Para cada par, determinar a diferença d_i, com sinal, entre os dois valores;

2. Atribuir postos a cada d_i, independentemente do sinal. No caso de d_i empatados, atribuir a média dos postos empatados;

3. Atribuir a cada posto o sinal + ou sinal – do d_i que ele representa;

4. Determinar *T* que é igual à menor das somas de postos de mesmo sinal;

5. Determinar *N* que é igual ao total de d_i com sinal;

6. Se $N \leq 25$, o Quadro 4 dá os valores críticos de *T* para diversos tamanhos de *N*. Se o valor observado de *T* não supera o valor indicado na tabela, para um nível de significância e um particular *N*, H_0 pode ser rejeitada.

N	Nível de significância para teste unilateral		
	0,025	0,01	0,005
	Nível de significância para teste unilateral		
	0,05	0,02	0,01
6	0		
7	2	0	
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Quadro 4 – Valores críticos do Testes de Wilcoxon

Se *N>25*, calcular valor de z através da fórmula:

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

Usando uma tabela estatística com a distribuição normal, determinar a probabilidade associada *p* sob H_0. No caso de uma prova bilateral, duplicar o valor de *p* dado. Se o *p* assim obtido não for superior ao nível de significância, rejeitar H_0.

4. Experiências comparativas

Neste Capítulo são apresentados os resultados experimentais dos vários modelos construídos, nomeadamente, para os modelos usando árvores de decisão e para os modelos usando *Random Forests*.

Todos os modelos foram construídos com base na mesma base de dados de informação e avaliados com as mesmas métricas, conforme seção 3.1.2, por forma a serem comparáveis.

De acordo com descrição efetuada na seção 3.1.3, podemos considerar que existem 5 abordagens diferentes:

Abordagem	Descrição
V0	Previsão com os parâmetros por defeito.
V1	Previsão usando pontos de corte previamente definidos para a probabilidade da classe APR.
V2	Previsão usando matriz de benefícios.
V3	Previsão usando matriz de benefícios e critério *expert* 1, baseado no descoberto provocado.
V4	Previsão usando matriz de benefícios e critério *expert* 2, baseado no descoberto provocado face ao montante da transação.

Quadro 5 – Abordagens do problema

Para cada abordagem, foram construídas 9 variantes do modelo de classificação usando o método das *Random Forests*.

A construção das 9 variantes resulta da utilização de diferentes valores para:

- o número máximo de árvores individuais a construir (ntree), tendo sido utilizados neste trabalho os valores 500, 1000 e 1500.
- o número de variáveis escolhidas aleatoriamente em cada nó (mtry), tendo sido utilizados neste trabalho os valores 5, 7 e 10.

O Quadro 6 define as 9 variantes e os respetivos parâmetros para as abordagens V0, V2, V3 e V4.

Variante	ntree	mtry
v1	500	5
v2	1000	5
v3	1500	5
v4	500	7
v5	1000	7
v6	1500	7
v7	500	10
v8	1000	10
v9	1500	10

Quadro 6 – Definição das 9 variantes para as abordagens V0, V2, V3 e V4 das *Random Forests*.

Para a abordagem V1, foram ainda testados cortes de 0.6, 0.8 e 0.95 para a probabilidade da classe APR necessária para que a decisão final seja APR, conduzindo assim a 27 variantes em vez de 9.

O Quadro 7 define as 27 variantes e os respetivos parâmetros para a abordagem V1.

Variante	Cutoff	ntree	mtry	Variante	Cutoff	ntree	mtry	Variante	Cutoff	ntree	mtry
v1	0.6	500	5	v10	0.8	500	5	v19	0.95	500	5
v2	0.6	1000	5	v11	0.8	1000	5	v20	0.95	1000	5
v3	0.6	1500	5	v12	0.8	1500	5	v21	0.95	1500	5
v4	0.6	500	7	v13	0.8	500	7	v22	0.95	500	7
v5	0.6	1000	7	v14	0.8	1000	7	v23	0.95	1000	7
v6	0.6	1500	7	v15	0.8	1500	7	v24	0.95	1500	7
v7	0.6	500	10	v16	0.8	500	10	v25	0.95	500	10
v8	0.6	1000	10	v17	0.8	1000	10	v26	0.95	1000	10
v9	0.6	1500	10	v18	0.8	1500	10	v27	0.95	1500	10

Quadro 7 – Definição das variantes para a abordagem V1 das *Random Forests*.

Cada variante construída pelo método de *Random Forests* para classificação é designada por “RandFor_Vx.vy”, onde *x* significa a abordagem e o *y* a variante resultante da combinação de parâmetros. Sempre que não haja problemas de ambiguidade, os nomes dos modelos serão abreviados para “Vx.vy”.

Foram também construídos modelos de árvores de decisão, com valores de 0, 0.5 e 1 para o parâmetro que controla o nível de poda da árvore (se), para servirem de referência para o tipo de modelos atualmente em utilização pelo banco. Dado o número elevado de modelos construídos para a abordagem V1, usando *Random Forests*, e atendendo que não se pretende uma comparação exaustiva entre os dois

métodos, não foram criados modelos usando árvores de decisão para a abordagem V1.

O Quadro 8 define as 3 variantes para cada abordagem das Arvores de Decisão e os respetivos parâmetros, com exceção da abordagem V1.

Variante	Se
v1	0
v2	0.5
v3	1

Quadro 8 – Definição das variantes para Árvores de Decisão.

Cada modelo associado às árvores de decisão individuais é designado por "Arvore_Vx.vy", onde *x* significa a abordagem efetuada e o *y* a variante resultante da combinação de parâmetros.

Cada variante construída foi testada 10 vezes, usando a técnica *holdout*, sendo o conjunto de teste formado por 30% dos dados. Os restantes 70% são utilizados na construção no modelo.

Do resumo dos 10 testes associados a cada tipo de modelo, são obtidas a média, o desvio padrão, o mínimo e o máximo de cada uma das métricas.

Por questões de confidencialidade, os resultados da métrica Lucro foram convertidos num índice, que consiste na razão entre o valor obtido para a métrica Lucro no modelo em análise e o valor médio mais elevado obtido do conjunto de todas as abordagens e variantes para a métrica Lucro.

Nas seções seguintes serão abordados os resultados mais relevantes para cada abordagem. No entanto, no Anexo B podem ser consultados todos os resultados obtidos para cada abordagem e para cada variante, com base nas estatísticas mencionadas.

4.1. Abordagem V0

Comece-se por analisar os vários modelos no que diz respeito à abordagem V0. Recorde-se que nesta abordagem a decisão é tomada com base apenas no modelo de previsão, sem utilização da matriz de benefícios.

No Quadro 9 a seguir apresentado constam os valores médios associados a cada métrica, para cada uma das variantes construídas.

Modelo	*Precision* APR	*Recall* APR	*Precision* REC	*Recall* REC	Erro	Lucro
V0.v1	0,9885	0,9983	0,9754	0,8517	0,0123	0,7406
V0.v2	0,9885	0,9983	0,9756	0,8518	0,0123	0,7382
V0.v3	0,9885	0,9984	0,9759	0,8520	0,0123	0,7385
V0.v4	0,9889	0,9982	0,9738	0,8564	0,0121	0,7420
V0.v5	0,9890	0,9983	0,9747	0,8580	0,0119	0,7425
V0.v6	0,9890	0,9983	0,9752	0,8574	0,0119	0,7420
V0.v7	0,9891	0,9982	0,9735	0,8599	0,0118	0,7441
V0.v8	0,9891	0,9982	0,9735	0,8593	0,0119	0,7432
V0.v9	0,9891	0,9982	0,9742	0,8593	0,0118	0,7436

Quadro 9 – Valores médios para cada métrica por cada variante da abordagem V0, usando *Random Forests*.

Relativamente à métrica Erro, os modelos com melhor média são os modelos V0.v7 e V0.v9.

Na Figura 11, usam-se *boxplots* para fornecer informação mais detalhada sobre a distribuição dos valores da métrica Erro nas 10 repetições, de que resultaram os valores médios mostrados no Quadro 9.

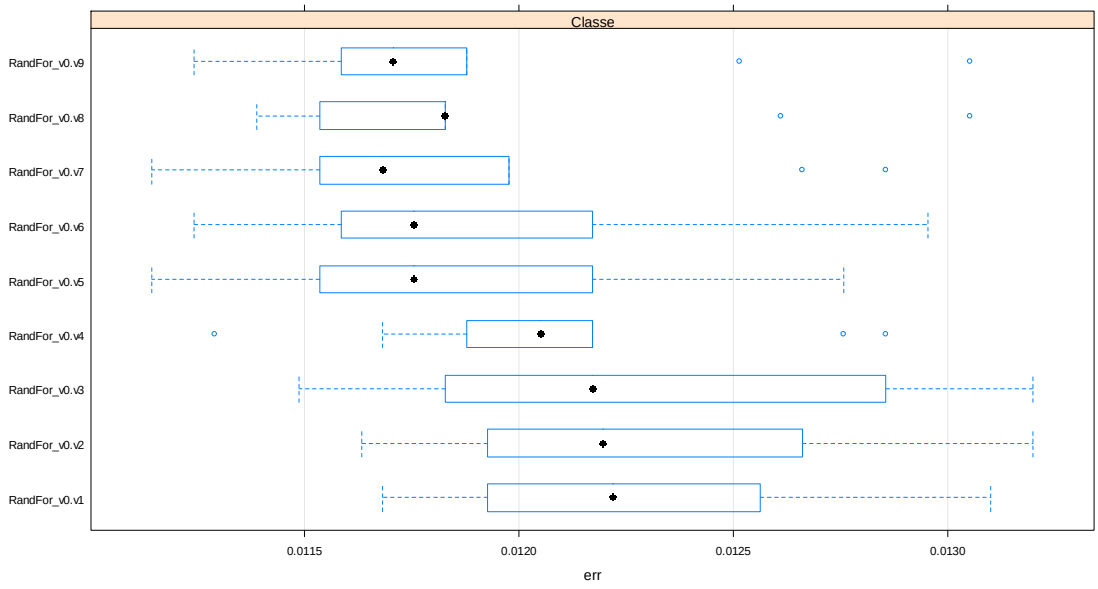


Figura 11 – Representação gráfica das estatísticas para a métrica Erro, usando *Random Forests*, para a abordagem V0.

Da informação contida na figura 11 podemos concluir que existe grande variabilidade nos resultados obtidos em cada uma das 10 repetições. Assim, importa avaliar, usando testes estatísticos de hipóteses, se as diferenças entre os valores médios da métrica Erro são estatisticamente significativas. Para este efeito foi usado o teste de *Wilcoxon* para amostras emparelhadas, que foi descrito na seção 3.3.2.. No Quadro 10 são mostrados os resultados de testes emparelhados, comparando o valor médio de um modelo base (escolheu-se os modelos V0.v7 e V0.v9, pois eram os que aparentavam ser os melhores), com os valores médios obtidos pelas restantes variantes.

	Modelos comparados								
Modelos base	**V0.v1**	**V0.v2**	**V0.v3**	**V0.v4**	**V0.v5**	**V0.v6**	**V0.v7**	**V0.v8**	**V0.v9**
V0.v7	+	+	+				n.a.		
V0.v9	+	+	+	+					n.a.

Quadro 10 – Significância estatística para a métrica Erro, usando *Random Forests*, para a abordagem V0.

Cada modelo, a que se chama modelo base, foi comparado com os restantes e o resultado da análise é:

- sinal '-': se o valor médio da métrica do modelo a comparar é significativamente inferior ao do modelo base, com um grau de confiança de 95%;
- sinal '--': se o valor médio da métrica do modelo a comparar é significativamente inferior ao do modelo base, com um grau de confiança de 99%;
- sinal '+': se o valor médio da métrica do modelo a comparar é significativamente superior ao do modelo base, com um grau de confiança de 95%;
- sinal '++': se o valor médio da métrica do modelo a comparar é significativamente superior ao do modelo base, com um grau de confiança de 99%;
- sem sinal: se não há significância estatística;
- n.a.: se não é aplicável.

De salientar que o significado de um valor médio significativamente inferior ou superior, pode ser diferente dependendo da métrica em análise. Assim, tratando-se de uma métrica a minimizar (por exemplo, Erro) interessa ter valores inferiores, enquanto que para métricas a maximizar (por exemplo, Lucro) já interessa ter valores superiores.

Pelo Quadro 10, pode-se constatar que o modelo V0.v7, no que concerne à métrica Erro, apresenta valores significativamente inferiores quando comparado com os modelos V0.v1, V0.v2 e V0.v3. Quanto ao modelo V0.v9, para a mesma métrica Erro, apresenta valores significativamente inferiores quando comparado com os modelos V0.v1, V0.v2, V0.v3 e V0.v4. Constata-se ainda que os modelos V0.v7 e V0.v9 não apresentam diferenças com significância estatística entre si, pelo que se poderia optar por qualquer um destes dois modelos, para a escolha de melhor modelo de acordo com a métrica Erro da abordagem V0.

No que diz respeito à métrica Lucro, o modelo que obtém o maior valor médio é o modelo V0.v7.

Na Figura 12, usam-se *boxplots* para fornecer informação mais detalhada sobre a distribuição dos valores da métrica Lucro nas 10 repetições, de que resultaram os valores médios mostrados no Quadro 9.

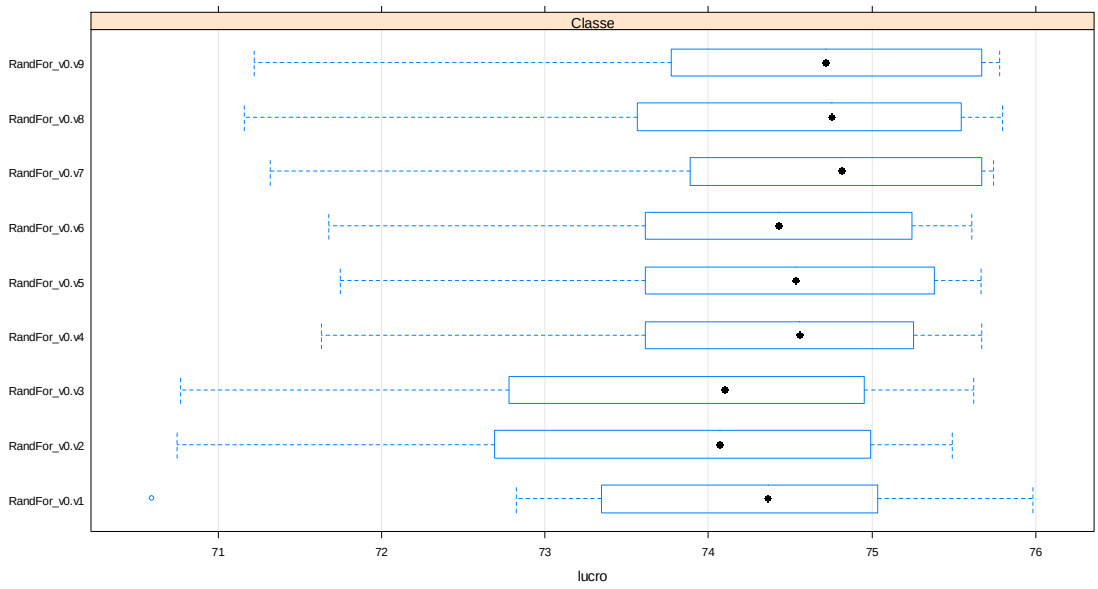


Figura 12 – Representação gráfica das estatísticas para a métrica Lucro, usando *Random Forests*, para a abordagem V0.

Recorrendo novamente à análise da significância estatística das diferenças observadas, no que diz respeito à métrica Lucro, obteve-se o Quadro 11 a seguir apresentado.

	Modelos comparados								
Modelo base	V0.v1	V0.v2	V0.v3	V0.v4	V0.v5	V0.v6	V0.v7	V0.v8	V0.v9
V0.v7	-	-	-	-			n.a.		

Quadro 11 – Significância estatística para a métrica Lucro, usando *Random Forests*, para a abordagem V0.

Recorde-se que o sinal "-" é utilizado quando o valor médio da métrica do modelo a comparar é significativamente inferior ao do modelo base, com um grau de confiança de 95%. Como o objetivo é maximizar o lucro, constata-se que a performance do modelo V0.v7 é melhor do que a dos modelos V0.v1, V0.v2, V0.v3 e V0.v4 e essa diferença é estatisticamente significativa, com um intervalo de confiança de 95%. Por outras palavras, o modelo V0.v7 construído usando uma seleção aleatória de 10 variáveis em cada nó obteve melhores resultados do que usando apenas 5 variáveis (modelos V0.v1, V0.v2 e V0.v3) e do que o modelo V0.v4 que usava 7 variáveis. Entre o modelo V0.v7 e os modelos V0.v5, V0.v6, V0.v8 e V0.v9 não se verificam diferenças com significância estatística.

Dado que a variante V0.v7 obteve uma média de 74,41% para a métrica Lucro e que este valor compara favoravelmente com as restantes variantes, se fosse necessário escolher a melhor variante da abordagem V0 para a métrica Lucro, escolher-se-ia a variante V0.v7.

Para comparação, a seguir é apresentada a Figura 13 e a Figura 14, com as *boxplots* para fornecer informação mais detalhada sobre a distribuição dos valores das métricas Lucro e Erro obtidos nas 10 repetições para cada variante das árvores de decisão, acompanhadas da *boxplot* da variante das *Random Forests* que obteve melhores resultados para a respetiva métrica.

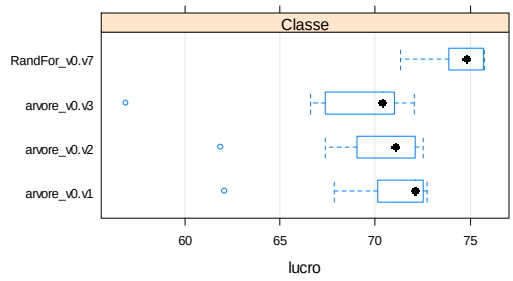


Figura 13 – *Boxplots* para a métrica Lucro, usando Árvores de Decisão, para a abordagem V0.

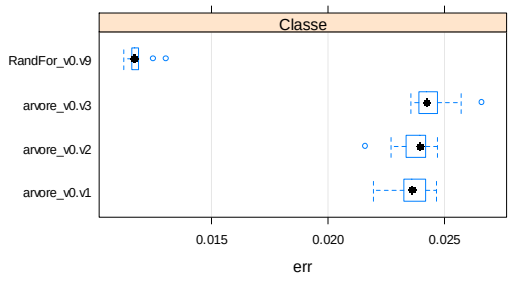


Figura 14 – *Boxplots* para a métrica Erro, usando Árvores de Decisão, para a abordagem V0.

Constata-se, pela Figura 13, que os valores obtidos para a métrica Lucro são inferiores aos obtidos pela variante RandFor_V0.v7. A variante arvore_V0.v1 é a que apresenta melhor resultado para a métrica Lucro, no que diz respeito a Árvores de Decisão. Na Figura 14, pode-se constatar que, para a métrica Erro, os valores das variantes das Árvores de Decisão são aproximadamente duas vezes superiores aos obtidos pela variante RandFor_V0.v9, para a mesma abordagem. A variante arvore_V0.v1 é também a variante que apresenta melhor resultado para a métrica Erro no grupo das Árvores de Decisão.

Note-se ainda que a dispersão dos resultados dos vários modelos para cada variante é mais elevada usando Árvores de Decisão do que *Random Forests*.

Recorrendo novamente à análise da significância estatística das diferenças observadas, para a variante com melhor resultado das *Random Forests* com as várias variantes das Árvores de Decisão, para as métricas Erro e Lucro, obteve-se o Quadro 12 e o Quadro 13.

	Modelos comparados		
Modelo base	Arvore_V0.v1	Arvore_V0.v2	Arvore_V0.v3
RandFor_V0.v9	+	+	+

Quadro 12 – Significância estatística para a métrica Erro, para a abordagem V0.

	Modelos comparados		
Modelo base	Arvore_V0.v1	Arvore_V0.v2	Arvore_V0.v3
RandFor_V0.v7	-	-	-

Quadro 13 – Significância estatística para a métrica Lucro, para a abordagem V0.

Pelo Quadro 12, pode-se constatar que todas as variantes das Árvores de Decisão, no que concerne à métrica Erro, apresentam valores significativamente superiores a variante RandFor_V0.v9.. Pelo Quadro 13, pode-se constatar que todas as variantes das Árvores de Decisão, no que concerne à métrica Lucro, apresentam valores significativamente inferiores a variante RandFor_V0.v7.

Quanto às métricas *Precision* e *Recall*, na Figura 15 constam as *boxplots* com a informação detalhada sobre a distribuição dos valores destas métricas, obtidos nas 10 repetições para cada variante das *Random Forests*.

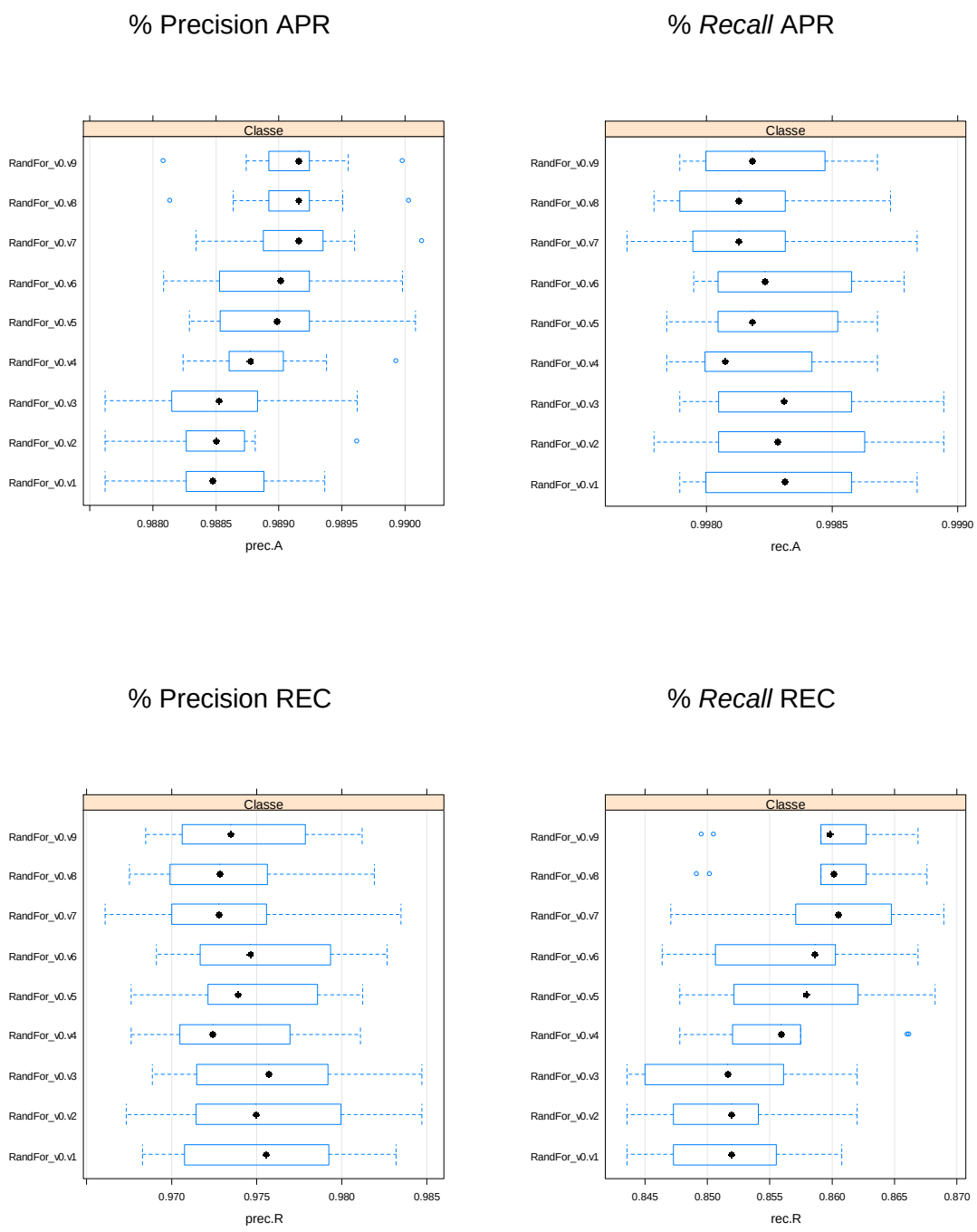


Figura 15 – *Boxplots* das métricas *Precision* e *Recall* para as classes APR e REC, usando *Random Forests*, para a abordagem V0.

Destas figuras observamos que, para todos os modelos desta abordagem, os valores da *Precision* para classe Aprovação (APR) situam-se sempre acima dos 98% e para a classe Recusa (REC) são sempre superiores a 96%. Para a métrica *Recall* associada à classe APR temos valores acima dos 99% e para a mesma métrica associada à classe REC assume valores acima de 84%. Esta última percentagem, com mais de 10 pontos percentuais abaixo das restantes, alerta que no total de transações que

deveriam ter decisão de recusa, cerca de 16% tiveram decisão de aprovação. A consequência é a criação de descobertos que não serão regularizados nos 22 dias úteis seguintes, que afetam o lucro obtido.

4.2. Abordagem V1

Quanto à abordagem V1, recorde-se que se pretende testar vários cortes no valor da probabilidade da classe APR necessária para que a decisão final seja APR.

No Quadro 14, constam os valores médios associados a cada métrica, para cada uma das variantes construídas.

Modelo	*Precision* APR	*Recall* APR	*Precision* REC	*Recall* REC	Erro	Lucro
V1.v1	0,9912	0,9965	0,9515	0,8874	0,0114	0,7578
V1.v2	0,9912	0,9965	0,9517	0,8869	0,0115	0,7579
V1.v3	0,9912	0,9965	0,9516	0,8872	0,0115	0,7577
V1.v4	0,9915	0,9964	0,9510	0,8910	0,0112	0,7604
V1.v5	0,9915	0,9964	0,9509	0,8902	0,0113	0,7598
V1.v6	0,9915	0,9965	0,9515	0,8905	0,0112	0,7601
V1.v7	0,9916	0,9963	0,9493	0,8923	0,0113	0,7614
V1.v8	0,9916	0,9963	0,9490	0,8916	0,0113	0,7611
V1.v9	0,9916	0,9963	0,9490	0,8920	0,0113	0,7613
V1.v10	0,9956	0,9841	0,8231	0,9438	0,0188	0,7831
V1.v11	0,9956	0,9843	0,8244	0,9442	0,0186	0,7836
V1.v12	0,9956	0,9842	0,8235	0,9441	0,0187	0,7834
V1.v13	0,9956	0,9840	0,8224	0,9444	0,0188	0,7837
V1.v14	0,9956	0,9842	0,8242	0,9444	0,0186	0,7830
V1.v15	0,9956	0,9843	0,8247	0,9443	0,0186	0,7831
V1.v16	0,9957	0,9839	0,8210	0,9451	0,0189	0,7827
V1.v17	0,9956	0,9840	0,8220	0,9448	0,0188	0,7827
V1.v18	0,9956	0,9840	0,8221	0,9449	0,0188	0,7827
V1.v19	0,9986	0,9031	0,4426	0,9842	0,0910	0,8187
V1.v20	0,9987	0,9029	0,4423	0,9846	0,0911	0,8188
V1.v21	0,9987	0,9027	0,4416	0,9844	0,0914	0,8188
V1.v22	0,9986	0,9060	0,4500	0,9841	0,0884	0,8205
V1.v23	0,9986	0,9055	0,4487	0,9843	0,0888	0,8202
V1.v24	0,9986	0,9057	0,4491	0,9841	0,0887	0,8191
V1.v25	0,9986	0,9086	0,4569	0,9836	0,0860	0,8207
V1.v26	0,9986	0,9081	0,4556	0,9840	0,0864	0,8208
V1.v27	0,9986	0,9080	0,4553	0,9840	0,0865	0,8208

Quadro 14 – Valores médios para cada métrica por cada variante da abordagem V1, usando *Random Forests*.

Relativamente à métrica Lucro, atendendo ao Quadro 14, constata-se que, para a abordagem V1, a variante que obtém a melhor média é a V1.v27, que regista o valor de 82,08%. Saliente-se que este valor é superior ao obtido nos modelos da abordagem V0, em que a classe atribuída em cada folha era aquela que registava maior proporção de elementos na folha, ou seja, o ponto de corte era 50%. No modelo V1.v27 define-se que apenas se considera a classe APR se a proporção de casos APR na folha for superior 95%.

A Figura 16 contém as *boxplots* com a informação detalhada sobre a distribuição dos valores da métrica Lucro, obtidos nas 10 repetições para cada variante das *Random Forests*.

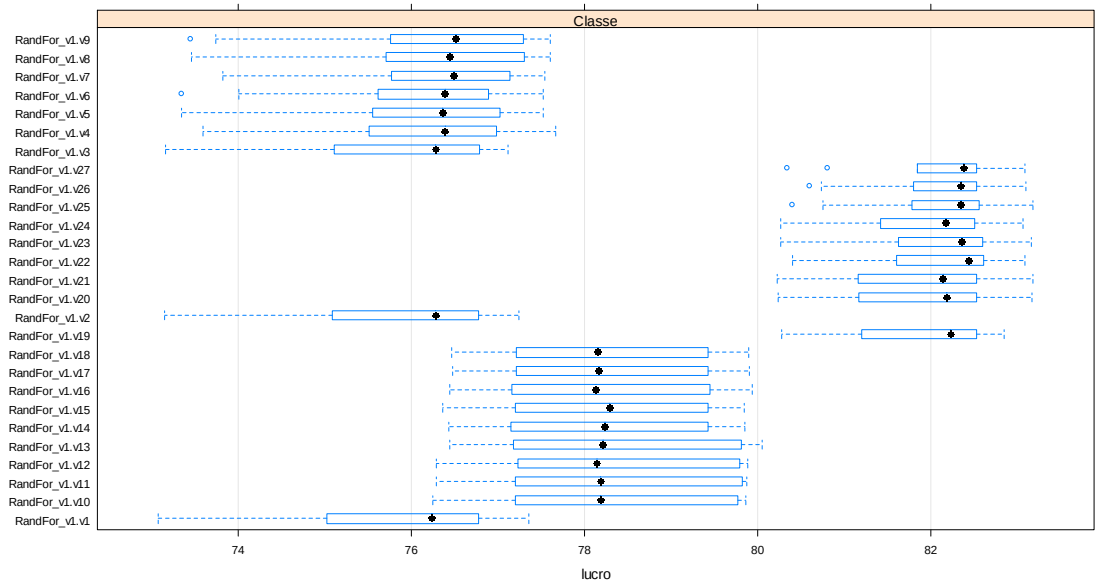


Figura 16 – *Boxplots* para a métrica Lucro, usando *Random Forests*, para a abordagem V1.

A Figura 16 evidencia que, é possível dividir a abordagem V1 em três grupos. Do primeiro grupo constam os modelos de V1.v1 até V1.v9. O segundo inclui os modelos de V1.v10 até V1.v18. Por fim, o terceiro grupo, inclui os modelos V1.v19 até V1.v27. Claramente, os modelos do terceiro grupo apresentam melhores resultados dos que os restantes, no que diz respeito à métrica Lucro.

Para comprovar que as diferenças entre os vários grupos são estatisticamente significativas, escolha-se a variante de cada grupo que registe o valor médio da métrica Lucro mais elevado, ou seja, escolha-se a variante V1.v7 do primeiro grupo, V1.v13 do segundo grupo e as variantes V1.v26.e V1.v27 do terceiro grupo, visto

registarem igual valor médio para a métrica Lucro. Os resultados obtidos encontram-se no Quadro 15 a seguir apresentado.

	Modelos comparados (1º Grupo)								
Modelos base	**V1.v1**	**V1.v2**	**V1.v3**	**V1.v4**	**V1.v5**	**V1.v6**	**V1.v7**	**V1.v8**	**V1.v9**
V1.v7	-	-	-		-	-	n.a.		
V1.v13	-	-	-	-	-	-	-	-	-
V1.v26	-	-	-	-	-	-	-	-	-
V1.v27	-	-	-	-	-	-	-	-	-

	Modelos comparados (2º Grupo)								
Modelos base	**V1.v10**	**V1.v11**	**V1.v12**	**V1.v13**	**V1.v14**	**V1.v15**	**V1.v16**	**V1.v17**	**V1.v18**
V1.v7	+	+	+	+	+	+	+	+	+
V1.v13				n.a.			-		
V1.v26	-	-	-	-	-	-	-	-	-
V1.v27	-	-	-	-	-	-	-	-	-

	Modelos comparados (3º Grupo)								
Modelos base	**V1.v19**	**V1.v20**	**V1.v21**	**V1.v22**	**V1.v23**	**V1.v24**	**V1.v25**	**V1.v26**	**V1.v27**
V1.v7	+	+	+	+	+	+	+	+	+
V1.v13	+	+	+	+	+	+	+	+	+
V1.v26								n.a.	
V1.v27	-					-			n.a.

Quadro 15 – Significância estatística para a métrica Lucro, usando *Random Forests* para a abordagem V1.

Alerte-se que valores negativos nas células dos quadros acima apresentados significam que o modelo comparado regista valores inferiores ao modelo base, podendo-se por isso afirmar que o modelo base obtém melhores resultados para a métrica Lucro. Assim sendo, no Quadro 15, constata-se que a variante V1.v7 apenas consegue registar valores superiores para a métrica Lucro, com significância estatística, quando comparados com modelos dentro do mesmo grupo que usam um menor número de variáveis para avaliar cada nó (variantes V1.v1, V1.v2, V1.v3, V1.v5 e V1.v6). A única exceção é a variante V1.v4, que é construída com o mesmo número de árvores e o mesmo corte que a variante V1.v7, diferindo apenas no número de variáveis avaliadas em cada nó, pois usa 7 em vez de 10. Estas semelhanças entre as duas variantes (V1.v7 e V1.v4) conduzem à inexistência de significância estatística para as diferenças obtidas nos resultados.

Pode-se constatar que a variante V1.v13, que é do segundo grupo, regista valores superiores e estatisticamente significativos, quando comparados com os modelos do primeiro grupo. Por outro lado, esta variante regista valores inferiores e estatisticamente significativos, quando comparada com os modelos do terceiro grupo. Não se verificam diferenças estatisticamente significativas, quando comparada com modelos do mesmo grupo.

Confirma-se ainda que as variantes V1.v26 e V1.v27, que pertencem ao terceiro grupo, registam valores superiores e estatisticamente significativos, quando comparados com os modelos do primeiro e do segundo grupo. A variante V1.v26 não verifica diferenças estatisticamente significativas com as variantes do mesmo grupo. Já a variante V1.v27 apresenta, com significância estatística, valores superiores à variante V1.19, pois esta última utiliza um menor número de árvores (500) e um menor número de variáveis (5) a testar em cada nó. A variante V1.v27 apresenta também valores superiores ao modelo e V1.v24, cuja única diferença entre as variantes é o número de variáveis a testar em cada nó, que para a variante V1.v27 são 10 e para a variante V1.v24 são 7.

Dado que a variante V1.v27 obteve uma média de 82,08% para a métrica Lucro e que este valor compara favoravelmente com as restantes variantes, se fosse necessário escolher a melhor variante da abordagem V1 para a métrica Lucro, escolher-se-ia a variante V1.v27.

Quanto à métrica Erro, através do Quadro 14, constata-se que a minimização desta métrica ocorre nos modelo V1.v4 e V1.v6, onde é atingido o valor de 0,0112.

Na Figura 17, constam as *boxplots* com a informação detalhada sobre a distribuição dos valores da métrica Erro, obtidos nas 10 repetições para cada variante das *Random Forests*.

Note-se, atendendo aos três grupos anteriormente definidos, que as variantes que apresentam melhores resultados para a métrica Erro (V1.v4 e V1.v6) estão incluídas no primeiro grupo, ou seja, o grupo que obtém piores valores para a métrica Lucro.

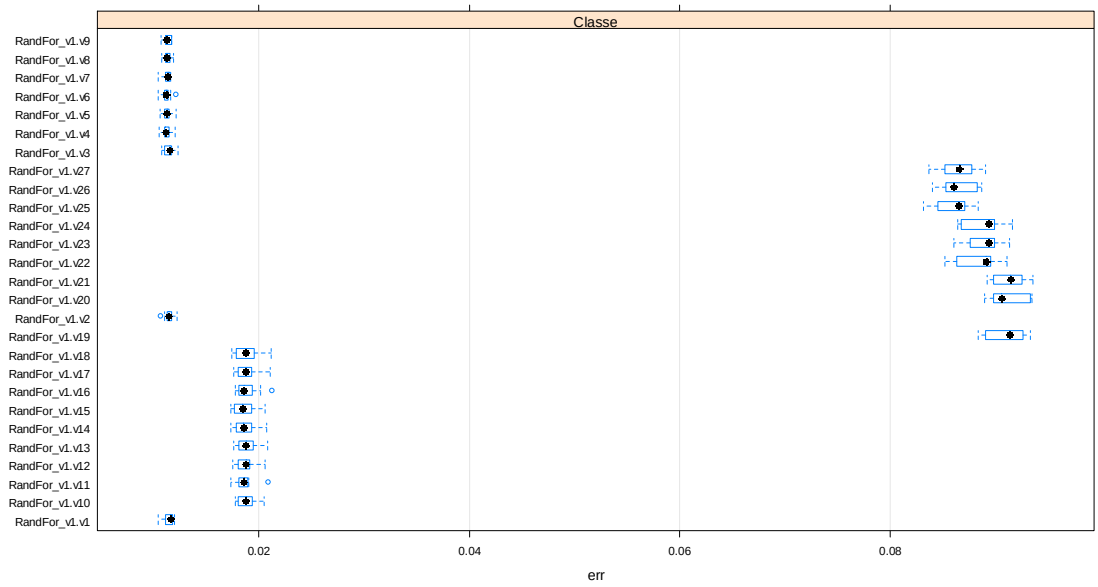


Figura 17 – *Boxplots* para a métrica Erro, usando *Random Forests*, para a abordagem V1.

Para confirmar se os modelos das variantes V1.v4 e V1.v6 apresentam diferenças estatisticamente significativas face aos restantes modelos, apresenta-se de seguida o Quadro 16, com o resultado dessa análise.

	Modelos comparados (1º Grupo)								
Modelos base	**V1.v1**	**V1.v2**	**V1.v3**	**V1.v4**	**V1.v5**	**V1.v6**	**V1.v7**	**V1.v8**	**V1.v9**
V1.v4	+	+	+	n.a.					
V1.v6		+	+			n.a.			

	Modelos comparados (2º Grupo)								
Modelos base	**V1.v10**	**V1.v11**	**V1.v12**	**V1.v13**	**V1.v14**	**V1.v15**	**V1.v16**	**V1.v17**	**V1.v18**
V1.v4	+	+	+	+	+	+	+	+	+
V1.v6	+	+	+	+	+	+	+	+	+

	Modelos comparados (3º Grupo)								
Modelos base	**V1.v19**	**V1.v20**	**V1.v21**	**V1.v22**	**V1.v23**	**V1.v24**	**V1.v25**	**V1.v26**	**V1.v27**
V1.v4	+	+	+	+	+	+	+	+	+
V1.v6	+	+	+	+	+	+	+	+	+

Quadro 16 – Significância estatística para a métrica Erro, usando *Random Forests*, para a abordagem V1

Alerte-se que valores positivos nas células dos quadros acima apresentados significam que o modelo comparado regista valores superiores ao modelo base,

podendo-se por isso afirmar que o modelo base obtém melhores resultados para a métrica Erro.

Dado que as variantes V1.v4 e V1.v6 obtiveram uma média de 0,0112 para a métrica Erro e que este valor compara favoravelmente com as restantes variantes, se fosse necessário escolher a melhor variante da abordagem V1 para a métrica Erro, escolher-se-ia qualquer uma destas variantes.

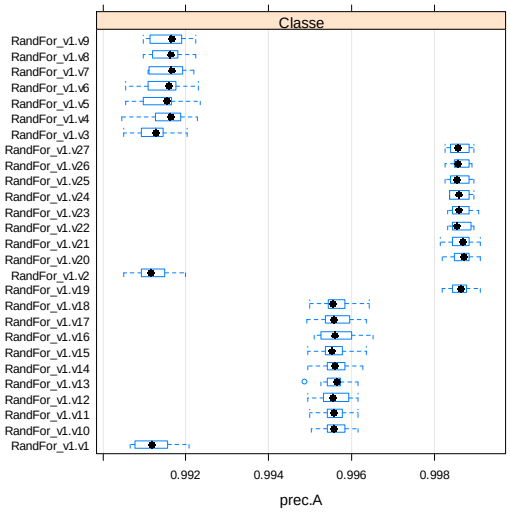
Quanto às métricas *Precision* e *Recall*, na Figura 18 constam as *boxplots* com a informação detalhada sobre a distribuição dos valores destas métricas, obtidos nas 10 repetições para cada variante das *Random Forests*.

No que diz respeito às métricas *Precision* e *Recall* para a classe REC, através da Figura 18, constata-se, que para o primeiro grupo, os valores da métrica *Precision* para a classe REC estão sempre acima dos 93% e os valores da métrica *Recall* para a classe REC estão sempre acima de 87%. Quando se observa o segundo grupo, consta-se que os valores da métrica *Precision* para a classe REC diminuem, podendo ser atingindo um mínimo de 80% e os valores da métrica *Recall* para a classe REC aumentam e estão sempre acima de 93%. Já no terceiro grupo, os valores da métrica *Precision* para a classe REC diminuem novamente, atingindo um mínimo de 42%, enquanto os valores da métrica *Recall* para a classe REC aumentam e estão sempre acima de 97%.

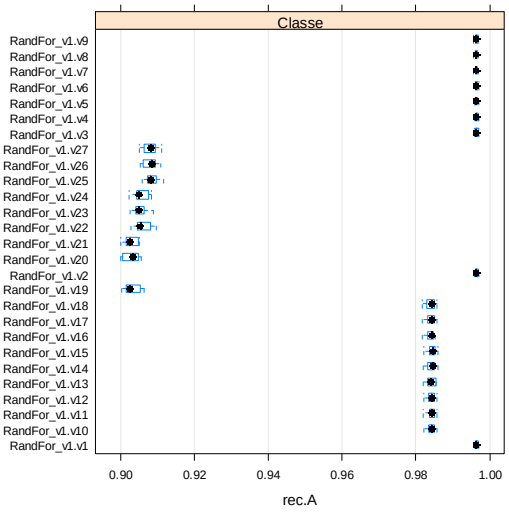
Quanto à métrica *Precision* para a classe APR regista valores acima dos 99%, com pequenas variações entre os três grupos. A métrica *Recall* para a classe APR regista, no primeiro grupo, valores acima de 99%, no segundo grupo, valores acima de 97% e no terceiro grupo, valores cima de 90%.

Estes resultados para as métricas Precison e *Recall* para a classe REC, aliados aos valores da métrica Lucro, justificam a abordagem V2, pois evidenciam que classificar com APR casos em que a classe deveria ser REC é mais penalizador para a métrica Lucro que atribuir classe REC quando a classe deveria ser APR.

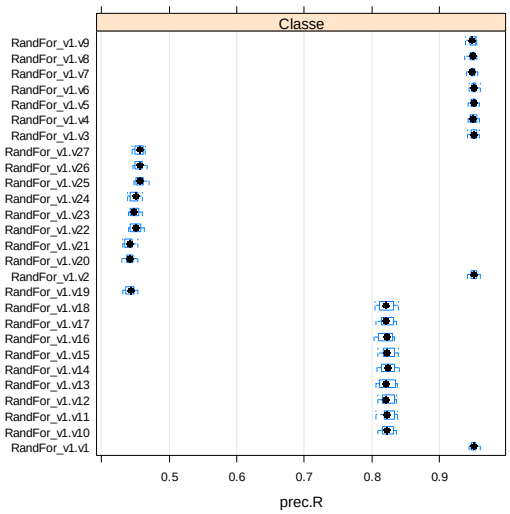
% Precision APR



% *Recall* APR



% Precision REC



% *Recall* REC

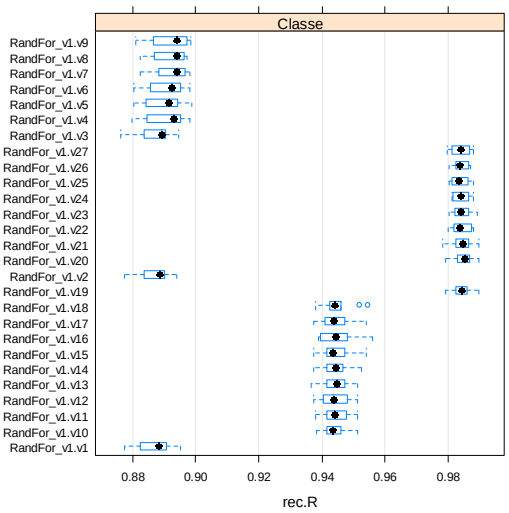


Figura 18 – *Boxplots* das métricas *Precision* e *Recall* para as classes APR e REC, usando *Random Forests*, para a abordagem V1.

4.3. Abordagem V2

Quanto à abordagem V2, procura-se melhorar os resultados através da utilização de uma matriz de benefícios.

No Quadro 17 a seguir apresentado constam os valores médios associados a cada métrica, para cada uma das variantes construídas.

Modelo	*Precision* APR	*Recall* APR	*Precision* REC	*Recall* REC	Erro	Lucro
V2.v1	0,9797	0,6103	0,1439	0,8384	0,3732	0,9941
V2.v2	0,9798	0,6095	0,1438	0,8390	0,3739	0,9954
V2.v3	0,9798	0,6095	0,1438	0,8390	0,3739	0,9954
V2.v4	0,9799	0,6149	0,1454	0,8384	0,3689	0,9918
V2.v5	0,9800	0,6144	0,1453	0,8392	0,3693	0,9926
V2.v6	0,9800	0,6143	0,1453	0,8393	0,3694	0,9928
V2.v7	0,9801	0,6188	0,1468	0,8393	0,3652	0,9897
V2.v8	0,9801	0,6181	0,1466	0,8394	0,3658	0,9906
V2.v9	0,9801	0,6179	0,1465	0,8395	0,3661	0,9909

Quadro 17 – Valores médios para cada métrica por cada variante da abordagem V2, usando *Random Forests*.

Por observação do Quadro 17, constata-se que, com esta nova abordagem, o valor mínimo obtido para a métrica Lucro foi de 98,97%, que ultrapassa claramente os resultados obtidos nas abordagens anteriores. A variante que apresenta o valor médio mais elevado para a métrica Lucro é a V2.v3, enquanto a variante V2.v7 é a que apresenta o valor médio mais reduzido para a métrica Erro.

Na Figura 19 e na Figura 20, pode-se observar as *boxplots* com a informação detalhada sobre a distribuição dos valores das métricas Lucro e Erro, obtidos nas 10 repetições para cada variante das *Random Forests*.

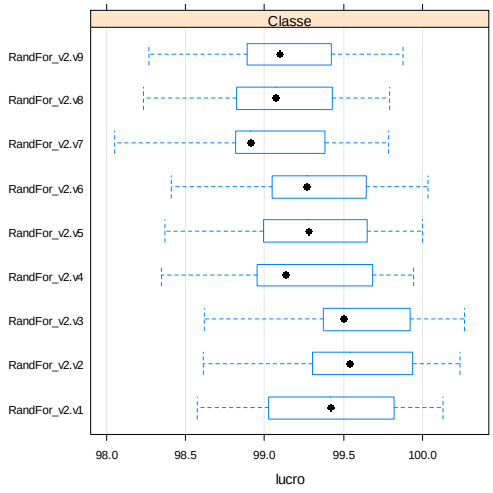


Figura 19 – *Boxplots* para a métrica Lucro, usando *Random Forests*, para a abordagem V2.

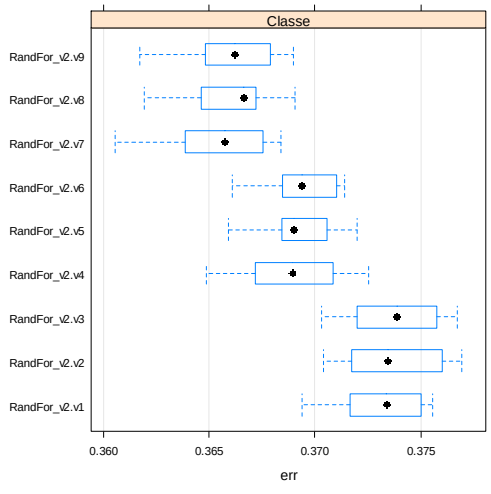


Figura 20 – *Boxplots* para a métrica Erro, usando *Random Forests*, para a abordagem V2.

Face às variações verificadas para cada variante, foi efetuada uma análise para validar se as diferenças entre as variantes são estatisticamente significativas, no que concerne à métrica Lucro. Os resultados obtidos constam do Quadro 18, apresentado a seguir.

	Modelos comparados								
Modelos base	**V2.v1**	**V2.v2**	**V2.v3**	**V2.v4**	**V2.v5**	**V2.v6**	**V2.v7**	**V2.v8**	**V2.v9**
V2.v3	-		n.a.	-	-	-	-	-	-

Quadro 18 – Significância estatística para a métrica Lucro, usando *Random Forests*, para a abordagem V2.

Em quase todas as comparações, as diferenças entre os modelos são estatisticamente significativas e as variantes com as quais a variante V2.v3 é comparada revelam valores inferiores. Assim sendo, considera-se a variante V2.v3 como a melhor variante da abordagem V2 para a métrica Lucro.

No que diz respeito à métrica Erro, o resultado dos testes de significância estatística constam do Quadro 19, a seguir apresentado.

	Modelos comparados								
Modelos base	**V2.v1**	**V2.v2**	**V2.v3**	**V2.v4**	**V2.v5**	**V2.v6**	**V2.v7**	**V2.v8**	**V2.v9**
V2.v7	+	+	+	+	+	+	n.a.		+

Quadro 19 – Significância estatística para a métrica Erro, usando *Random Forests*, para a abordagem V2.

Mais uma vez, podemos constatar que em quase todas as comparações, as diferenças entre os modelos são estatisticamente significativas e as variantes com as quais a variante V2.v7 é comparada revelam valores superiores. Assim sendo, considera-se a variante V2.v7 como a melhor variante da abordagem V2 para a métrica Erro.

Para comparação, a seguir é apresentada a Figura 21 e a Figura 22, com as *boxplots* com a informação detalhada sobre a distribuição dos valores das métricas Lucro e Erro, obtidos nas 10 repetições para cada variante das Árvores de Decisão, acompanhadas da *boxplot* da variante das *Random Forests* que obteve melhores resultados para a respetiva métrica.

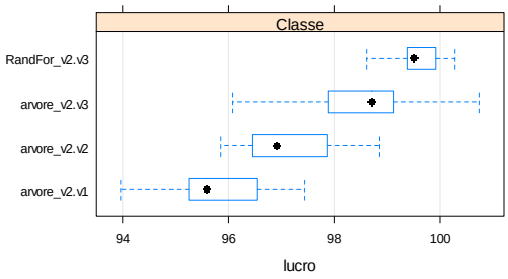


Figura 21 – *Boxplots* para a métrica Lucro, usando Árvores de Decisão, para a abordagem V2.

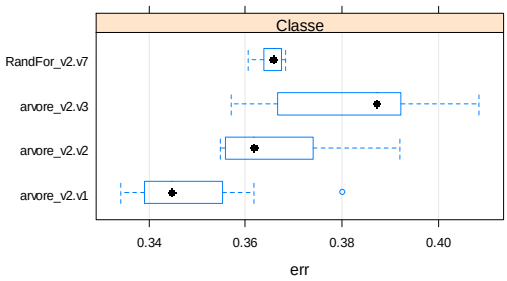


Figura 22 – *Boxplots* para a métrica Erro, usando Árvores de Decisão para a abordagem V2.

Constata-se, pela Figura 21, que os valores obtidos para a métrica Lucro são regra geral inferiores ao obtido pela variante RandFor_v2.v3 das Random Forest, para a abordagem V2.

Para a métrica Erro, na Figura 22, pode-se constatar que os valores desta métrica têm grande variabilidade entre as variantes, atingindo valores superiores e inferiores aos valores obtidos pela variante RandFor_v2.v7 das *Random Forests*. A variante arvore_v2.v1 é a que obtém valores inferiores aos valores da variante RandFor_v2.v7 das *Random Forests*. Porém, observando a Figura 21, é também esta variante que tem pior valor para a métrica Lucro. Inversamente, a variante arvore_v2.v3 é a que obtém piores valores para a métrica Erro, no entanto é que regista melhor valor para a métrica Lucro.

Note-se ainda que a dispersão dos resultados dos vários modelos para cada variante é mais elevada usando Árvores de Decisão do que *Random Forests*.

Recorrendo novamente à análise da significância estatística das diferenças observadas, para a variante com melhor resultado das *Random Forests* com as várias variantes das Árvores de Decisão, para as métricas Erro e Lucro, obteve-se o Quadro 20 e o Quadro 21.

Pelo Quadro 20 e pelo Quadro 21 comprova-se que as diferenças que levaram aos comentários anteriores têm significância estatística.

	Modelos comparados		
Modelo base	Arvore_V2.v1	Arvore_V2.v2	Arvore_V2.v3
RandFor_V2.v7	-		+

Quadro 20 – Significância estatística para a métrica Erro, para a abordagem V2.

	Modelos comparados		
Modelo base	Arvore_V2.v1	Arvore_V2.v2	Arvore_V2.v3
RandFor_V2.v3	-	-	-

Quadro 21 – Significância estatística para a métrica Lucro, para a abordagem V2.

Quanto às métricas *Precision* e *Recall*, na Figura 23 constam as *boxplots* com a informação detalhada sobre a distribuição dos valores destas métricas, obtidos nas 10 repetições para cada variante das *Random Forests*.

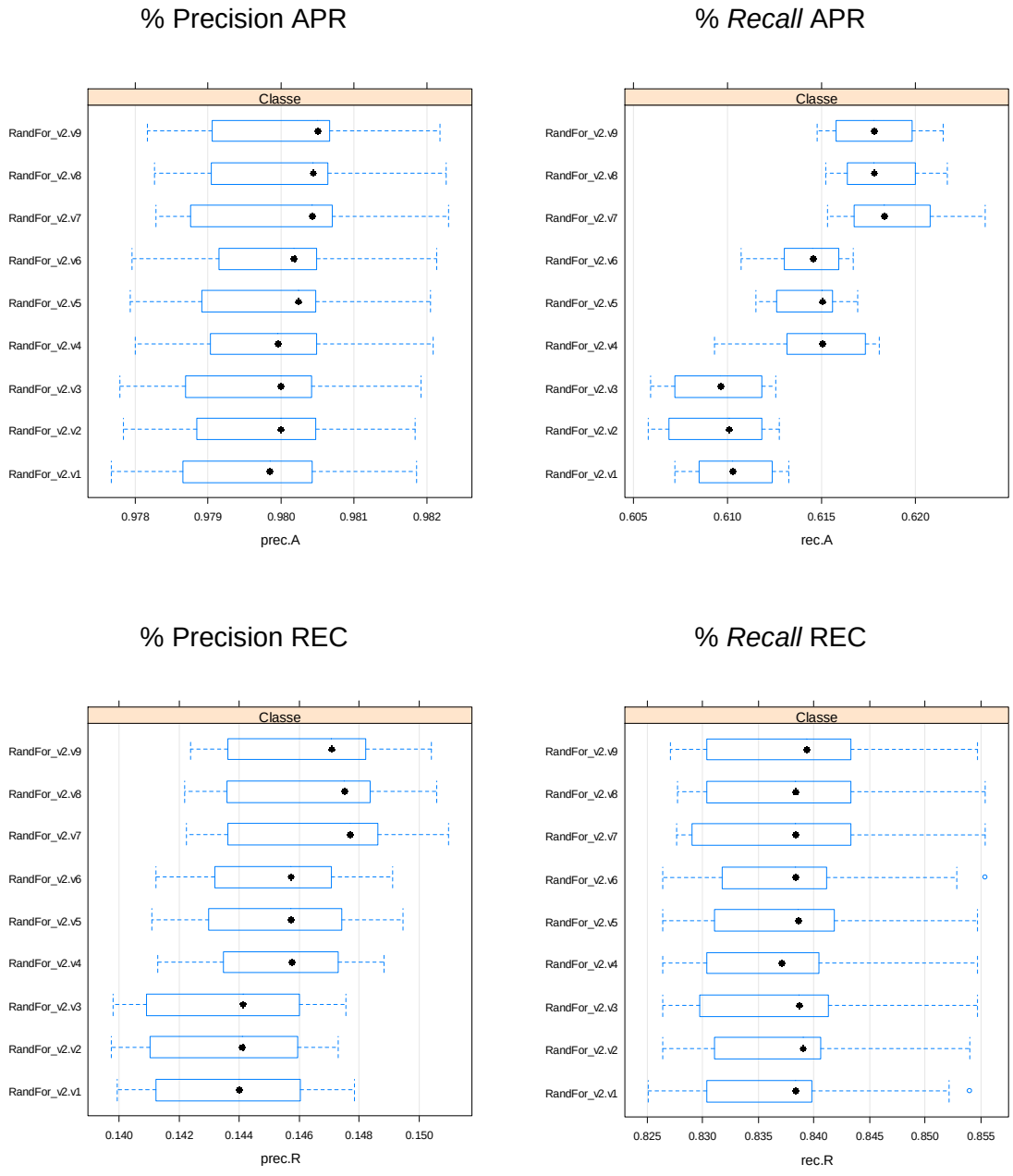


Figura 23 – *Boxplots* das métricas *Precision* e *Recall* para as classes APR e REC, usando *Random Forests*, para a abordagem V2.

Através da Figura 23, constata-se que os valores da métrica *Precision* para a classe REC estão sempre acima dos 13% e os valores da métrica *Recall* para a classe REC estão sempre acima de 82%, para cada uma das variantes.

A métrica *Precision* para a classe APR regista valores acima dos 97% e a métrica *Recall* para a classe APR regista valores acima de 60%, para cada uma das variantes.

Os resultados para as métricas Precison e *Recall*, evidenciam que existem casos extremos, que são erradamente classificados pelos modelos. Justifica-se assim a abordagem V3, que procura identificar e tratar esses casos extremos.

4.4. Abordagem V3

Apesar de com a abordagem V2 terem sido obtidos valores bastantes superiores às abordagens anteriores, para a métrica Lucro, procura-se na abordagem V3 melhorar esses resultados, através da inclusão de critérios *expert* baseado no descoberto provocado pela aprovação da transação.

No Quadro 22 a seguir apresentado constam os valores médios associados a cada métrica, para cada uma das variantes construídas.

Modelo	*Precision* APR	*Recall* APR	*Precision* REC	*Recall* REC	Erro	Lucro
V3.v1	0,9794	0,5988	0,1405	0,8389	0,3838	0,9990
V3.v2	0,9795	0,5983	0,1404	0,8395	0,3842	1,0000
V3.v3	0,9795	0,5984	0,1404	0,8395	0,3841	1,0000
V3.v4	0,9796	0,6034	0,1418	0,8389	0,3796	0,9968
V3.v5	0,9796	0,6030	0,1419	0,8397	0,3798	0,9974
V3.v6	0,9797	0,6031	0,1419	0,8397	0,3797	0,9974
V3.v7	0,9798	0,6073	0,1432	0,8398	0,3759	0,9946
V3.v8	0,9798	0,6068	0,1430	0,8399	0,3763	0,9954
V3.v9	0,9798	0,6066	0,1430	0,8399	0,3765	0,9957

Quadro 22 – Valores médios para cada métrica por cada variante da abordagem V3, usando *Random Forests*.

Por observação do Quadro 22, constata-se que com esta nova abordagem o maior valor médio obtido para a métrica Lucro é o das variantes V3.v2 e V3.v3, com o valor de 100%, que melhora um pouco os resultados obtidos nas abordagens anteriores. Por outro lado, o modelo V3.v7 é o que minimiza a métrica Erro, obtendo um valor de 0,3759. Constata-se ainda que a maximização do lucro não significa uma diminuição do Erro.

Na Figura 24 e na Figura 25, pode-se observar as *boxplots* com a informação detalhada sobre a distribuição dos valores das métricas Lucro e Erro, obtidos nas 10 repetições para cada variante das *Random Forests*.

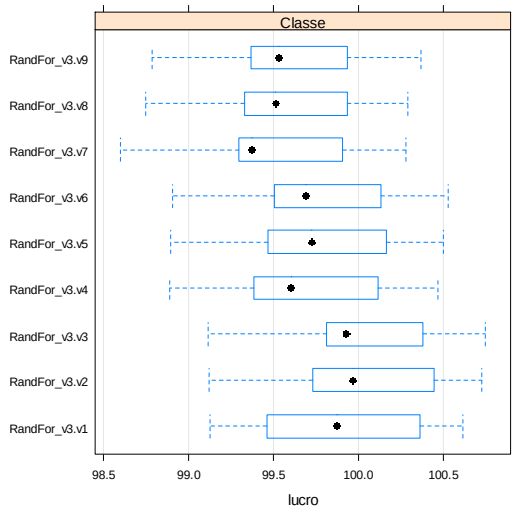


Figura 24 – *Boxplots* para a métrica Lucro, usando *Random Forests*, para a abordagem V3.

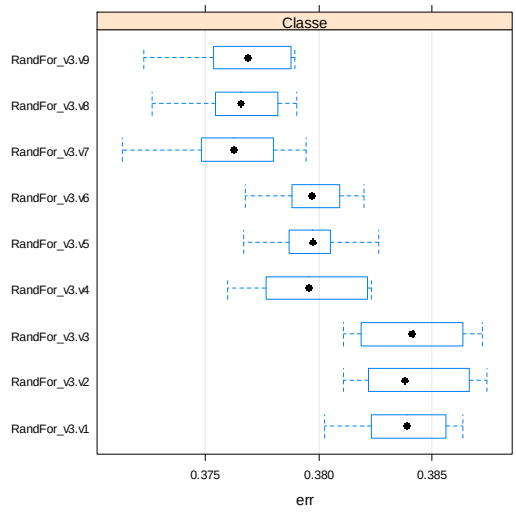


Figura 25 – *Boxplots* para a métrica Erro, usando *Random Forests*, para a abordagem V3.

Face às variações verificadas para cada variante, foi efetuada uma análise para validar se as diferenças entre as variantes são estatisticamente significativas, no que concerne à métrica Lucro. Os resultados obtidos constam do Quadro 23.

	Modelos comparados								
Modelos base	**V3.v1**	**V3.v2**	**V3.v3**	**V3.v4**	**V3.v5**	**V3.v6**	**V3.v7**	**V3.v8**	**V3.v9**
V3.v2	-	n.a.		-	-	-	-	-	-
V3.v3	-		n.a.	-	-	-	-	-	-

Quadro 23 – Significância estatística para a métrica Lucro, usando *Random Forests*, para a abordagem V3.

Em quase todas as comparações, as diferenças entre os modelos são estatisticamente significativas e as variantes com as quais as variantes V3.v2 e V3.v3 são comparadas revelam valores inferiores. Assim sendo, consideram-se estas variantes (V3.v2 e V3.v3) como as melhores variantes da abordagem V3 para a métrica Lucro.

No que diz respeito à métrica Erro, o resultado dos testes de significância estatística constam do Quadro 24, a seguir apresentado.

	Modelos comparados								
Modelos base	**V3.v1**	**V3.v2**	**V3.v3**	**V3.v4**	**V3.v5**	**V3.v6**	**V3.v7**	**V3.v8**	**V3.v9**
V3.v7	+	+	+	+	+	+	n.a.		+

Quadro 24 – Significância estatística para a métrica Erro, usando *Random Forests*, para a abordagem V3.

Mais uma vez, podemos constatar que em quase todas as comparações, as diferenças entre os modelos são estatisticamente significativas e as variantes com as quais a variante V3.v7 é comparada revelam valores superiores. Assim sendo, considera-se a variante V3.v7 como a melhor variante da abordagem V3 para a métrica Erro.

Para comparação, a seguir é apresentada a Figura 26 e a Figura 27, com as *boxplots* com a informação detalhada sobre a distribuição dos valores das métricas Lucro e Erro, obtidos nas 10 repetições para cada variante das Árvores de Decisão, acompanhadas da *boxplot* da variante das *Random Forests* que obteve melhores resultados para a respetiva métrica.

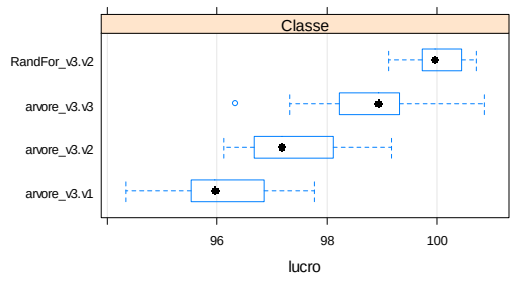


Figura 26 – *Boxplots* para a métrica Lucro, usando Árvores de Decisão, para a abordagem V3.

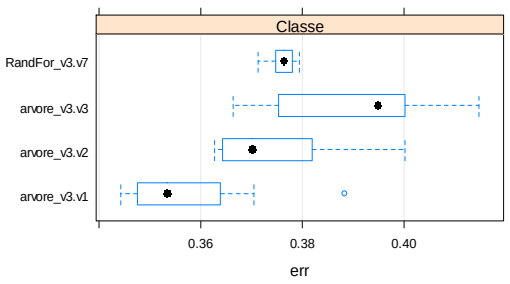


Figura 27 – *Boxplots* para a métrica Erro, usando Árvores de Decisão para a abordagem V3.

Constata-se, pela Figura 26, que os valores obtidos para a métrica Lucro, são em geral, inferiores aos obtidos pela variante RanFor_v3.v2 das *Random Forests*, para a abordagem V3. A variante arvore_v3.v3 é a que regista melhor valor para a métrica Lucro, contudo é também esta a variante que obtém os piores valores para a métrica Erro, conforme Figura 27. Observa-se ainda que, para a métrica Erro, os valores têm grande variabilidade entre as variantes, atingindo valores superiores e inferiores aos valores obtidos pela variante RandFor_v3.v7. A variante arvore_v3.v1 é a que regista um melhor valor para a métrica Erro, cuja mediana é inferior à obtida pelo método das *Random Forests*. Porém este modelo é também aquele que tem pior valor para a métrica Lucro.

Note-se ainda que a dispersão dos resultados dos vários modelos para cada variante é mais elevada usando Árvores de Decisão do que *Random Forests*.

Recorrendo novamente à análise da significância estatística das diferenças observadas, para a variante com melhor resultado das *Random Forests* com as várias variantes das Árvores de Decisão, para as métricas Erro e Lucro, obteve-se o Quadro 25 e o Quadro 26.

	Modelos comparados		
Modelo base	Arvore_V3.v1	Arvore_V3.v2	Arvore_V3.v3
RandFor_V3.v7	-		+

Quadro 25 – Significância estatística para a métrica Erro, para a abordagem V3.

	Modelos comparados		
Modelo base	Arvore_V3.v1	Arvore_V3.v2	Arvore_V3.v3
RandFor_V3.v2	-	-	-

Quadro 26 – Significância estatística para a métrica Lucro, para a abordagem V3.

Pelo Quadro 25 e o pelo Quadro 26 comprava-se que as diferenças que levaram aos comentários anteriores têm significância estatística.

Quanto às métricas *Precision* e *Recall*, na Figura 28 constam as *boxplots* com a informação detalhada sobre a distribuição dos valores destas métricas, obtidos nas 10 repetições para cada variante das *Random Forests*.

Através da Figura 28, constata-se que os valores da métrica *Precision* para a classe REC estão sempre acima dos 13% e os valores da métrica *Recall* para a classe REC estão sempre acima de 82%, para cada uma das variantes.

A métrica *Precision* para a classe APR regista valores acima dos 97% e a métrica *Recall* para a classe APR regista valores acima de 59%, para cada uma das variantes.

Apesar de com esta abordagem V3 terem sido obtidos valores superiores às abordagens anteriores, para a métrica Lucro, constata-se um aumento dos valores associados á métrica Erro, assim como uma ligeira redução nas métricas de *Precision* e *Recall* associadas a cada classe.

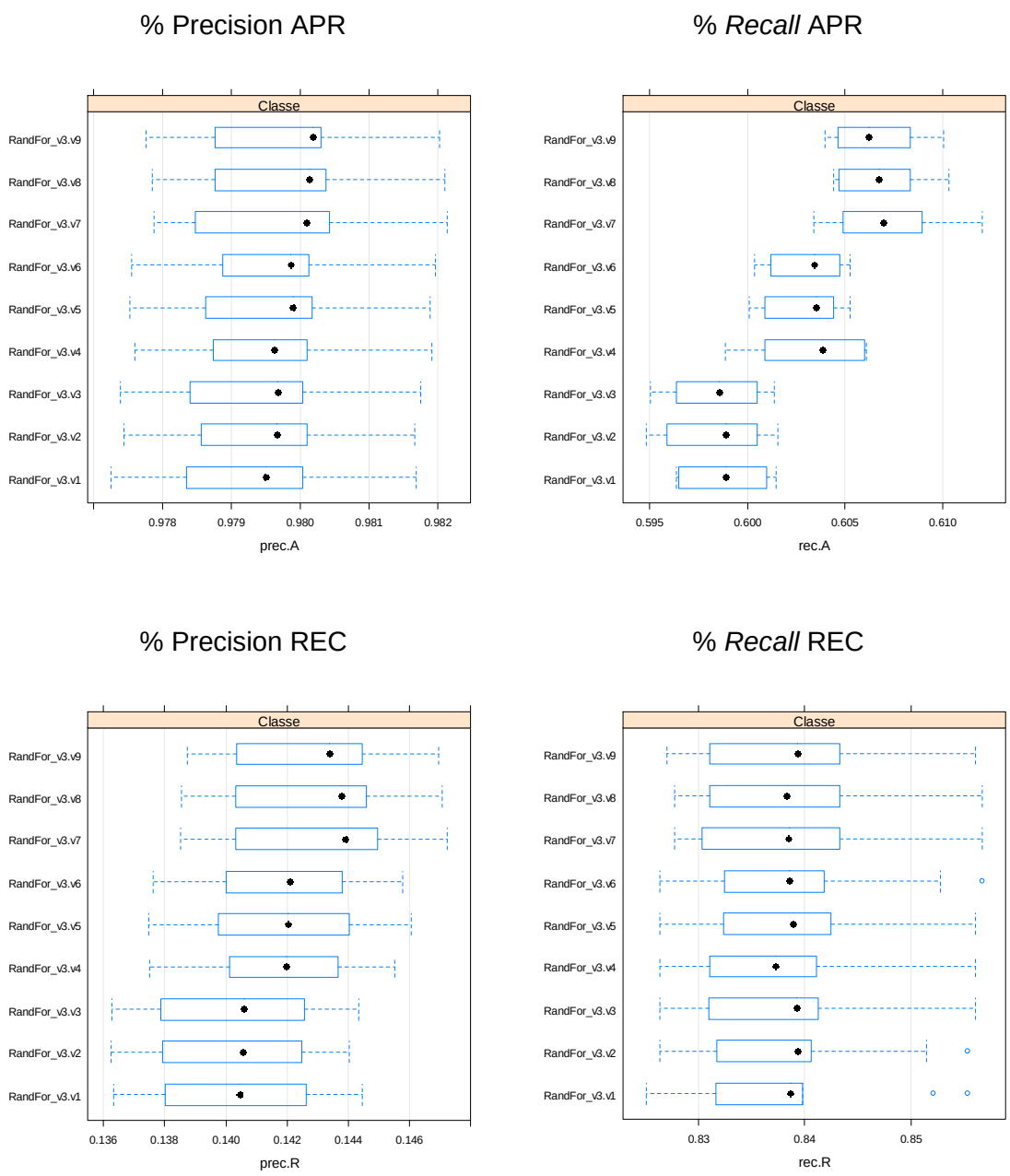


Figura 28 – *Boxplots* das métricas *Precision* e *Recall* para as classes APR e REC, usando *Random Forests*, para a abordagem V3.

4.5. Abordagem V4

Apesar de com a abordagem V3 terem sido obtidos valores superiores às abordagens anteriores, para a métrica Lucro, procura-se na abordagem V4 melhorar esses resultados, através da inclusão de um novo critério *expert* que relaciona o descoberto provocado com o montante da transação.

No Quadro 27 a seguir apresentado constam os valores médios associados a cada métrica, para cada uma das variantes construídas.

Modelo	*Precision* APR	*Recall* APR	*Precision* REC	*Recall* REC	Erro	Lucro
V4.v1	0,9805	0,6585	0,1600	0,8324	0,3289	0,9694
V4.v2	0,9805	0,6575	0,1597	0,8330	0,3298	0,9705
V4.v3	0,9805	0,6574	0,1597	0,8331	0,3299	0,9706
V4.v4	0,9807	0,6629	0,1618	0,8327	0,3248	0,9673
V4.v5	0,9807	0,6622	0,1616	0,8332	0,3254	0,9680
V4.v6	0,9807	0,6620	0,1616	0,8334	0,3255	0,9682
V4.v7	0,9808	0,6669	0,1636	0,8333	0,3210	0,9650
V4.v8	0,9808	0,6660	0,1632	0,8334	0,3219	0,9660
V4.v9	0,9808	0,6656	0,1630	0,8334	0,3222	0,9663

Quadro 27 – Valores médios para cada métrica por cada variante da abordagem V4, usando *Random Forests*.

Por observação do Quadro 27, constata-se que com esta nova abordagem o maior valor médio obtido para a métrica Lucro é o da variante V4.v3, com o valor de 97,06%. Por outro lado, a variante V4.v7 é a que minimiza a métrica Erro, obtendo um valor de 0,3210. Constata-se ainda que a maximização do lucro não significa uma diminuição do erro.

Note-se que para a métrica Lucro, os valores obtidos nesta abordagem são piores do que os valores obtidos nas abordagens V2 e V3. Porém são melhores do que os valores obtidos nas abordagens V0 e V1. Para a métrica Erro, os valores obtidos nesta abordagem são piores do que os valores obtidos nas abordagens V0 e V1. Porém são melhores do que os valores obtidos nas abordagens V2 e V3.

Na Figura 29 e na Figura 30, pode-se observar as *boxplots* com a informação detalhada sobre a distribuição dos valores das métricas Lucro e Erro, obtidos nas 10 repetições para cada variante das *Random Forests*.

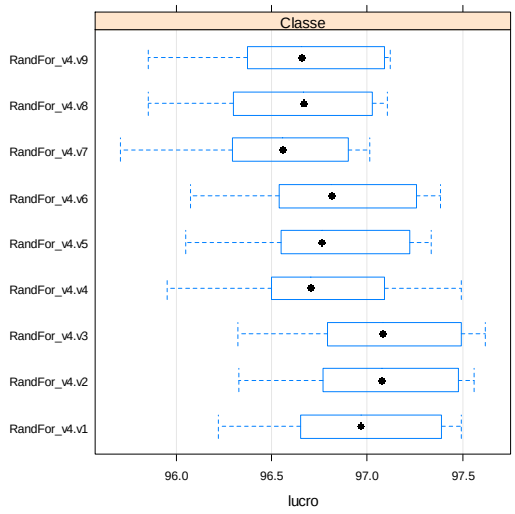


Figura 29 – *Boxplots* para a métrica Lucro, usando *Random Forests*, para a abordagem V4.

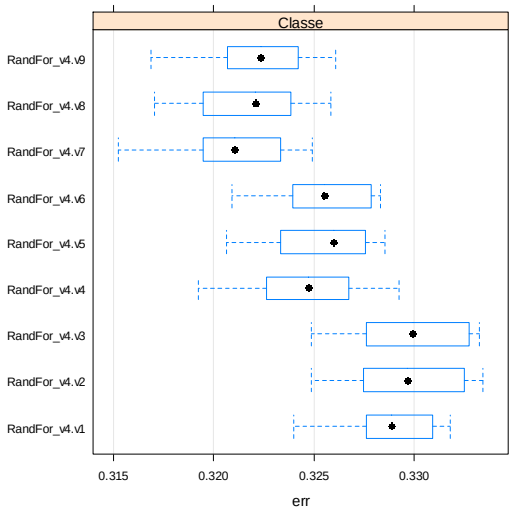


Figura 30 – *Boxplots* para a métrica Erro, usando *Random Forests*, para a abordagem V4.

Face às variações verificadas para cada variante, foi efetuada uma análise para validar se as diferenças entre as variantes são estatisticamente significativas, no que concerne à métrica Lucro. Os resultados obtidos constam do Quadro 28.

	Modelos comparados								
Modelos base	**V4.v1**	**V4.v2**	**V4.v3**	**V4.v4**	**V4.v5**	**V4.v6**	**V4.v7**	**V4.v8**	**V4.v9**
V4.v3	-		n.a.	-	-	-	-	-	-

Quadro 28 – Significância estatística para a métrica Lucro, usando *Random Forests*, para a abordagem V4.

Em quase todas as comparações, as diferenças entre os modelos são estatisticamente significativas e as variantes com as quais a variante V4.v3 é comparada revelam valores inferiores. Assim sendo, considera-se esta variante V4.v3 como a melhor variante da abordagem V4 para a métrica Lucro.

No que diz respeito à métrica Erro, o resultado dos testes de significância estatística constam do Quadro 29, a seguir apresentado.

	Modelos comparados								
Modelos base	**V4.v1**	**V4.v2**	**V4.v3**	**V4.v4**	**V4.v5**	**V4.v6**	**V4.v7**	**V4.v8**	**V4.v9**
V4.v7	+	+	+	+	+	+	n.a.	+	+

Quadro 29 – Significância estatística para a métrica Erro, usando *Random Forests*, para a abordagem V4.

No Quadro 29 podemos constatar que em todas as comparações, as diferenças entre os modelos são estatisticamente significativas e as variantes com as quais a variante V4.v7 é comparada revelam valores superiores. Assim sendo, considera-se a variante V4.v7 como a melhor variante da abordagem V4 para a métrica Erro.

Para comparação, a seguir é apresentada a Figura 31 e a Figura 32, com as *boxplots* com a informação detalhada sobre a distribuição dos valores das métricas Lucro e Erro, obtidos nas 10 repetições para cada variante das Árvores de Decisão, acompanhadas da *boxplot* da variante das *Random Forests* que obteve melhores resultados para a respetiva métrica.

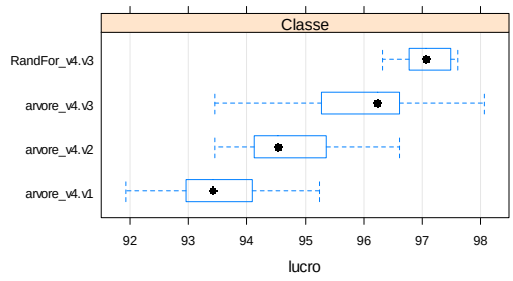


Figura 31 – *Boxplots* para a métrica Lucro, usando Árvores de Decisão, para a abordagem V4.

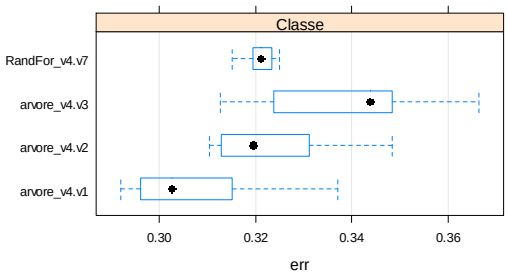


Figura 32 – *Boxplots* para a métrica Erro, usando Árvores de Decisão para a abordagem V4.

Constata-se, pela Figura 31, que os valores obtidos para a métrica Lucro, são em geral inferiores aos obtidos para a variante RandFor_V4.v3 das *Random Forests*, para a abordagem V4. De todas as variantes usando Árvores de Decisão, a variante arvore_v4.v3 é a que regista melhor valor para a métrica Lucro, contudo é também esta a variante que obtém os piores valores para a métrica Erro, conforme Figura 32. Observa-se ainda que, para a métrica Erro, os valores têm grande variabilidade entre as variantes. Do conjunto de modelos obtidos usando Árvores de Decisão, a variante arvore_v4.v1 é a que regista um melhor valor para a métrica Erro, cuja mediana chega a ser inferior à obtida pela variante RandFor_V4.v7 das *Random Forests*. Porém este modelo é também aquele que tem pior valor para a métrica Lucro.

Note-se mais uma vez, que a dispersão dos resultados dos vários modelos para cada variante é mais elevada usando Árvores de Decisão do que *Random Forests*.

Recorrendo novamente à análise da significância estatística das diferenças observadas, para a variante com melhor resultado das *Random Forests* com as várias

variantes das Árvores de Decisão, para as métricas Erro e Lucro, obteve-se o Quadro 30 e o Quadro 31.

	Modelos comparados		
Modelo base	Arvore_V4.v1	Arvore_V4.v2	Arvore_V4.v3
RandFor_V4.v7	-		+

Quadro 30 – Significância estatística para a métrica Erro, para a abordagem V4.

	Modelos comparados		
Modelo base	Arvore_V4.v1	Arvore_V4.v2	Arvore_V4.v3
RandFor_V4.v3	-	-	-

Quadro 31 – Significância estatística para a métrica Lucro, para a abordagem V4.

Pelo Quadro 30 e pelo Quadro 31 comprova-se que as diferenças que levaram aos comentários anteriores têm significância estatística.

Quanto às métricas *Precision* e *Recall*, na Figura 33 constam as *boxplots* com a informação detalhada sobre a distribuição dos valores destas métricas, obtidos nas 10 repetições para cada variante das *Random Forests*.

Através da Figura 33, constata-se que os valores da métrica *Precision* para a classe REC estão sempre acima dos 15% e os valores da métrica *Recall* para a classe REC estão sempre acima de 81%, para cada uma das variantes.

A métrica *Precision* para a classe APR regista valores acima dos 97% e a métrica *Recall* para a classe APR regista valores acima de 65%, para cada uma das variantes.

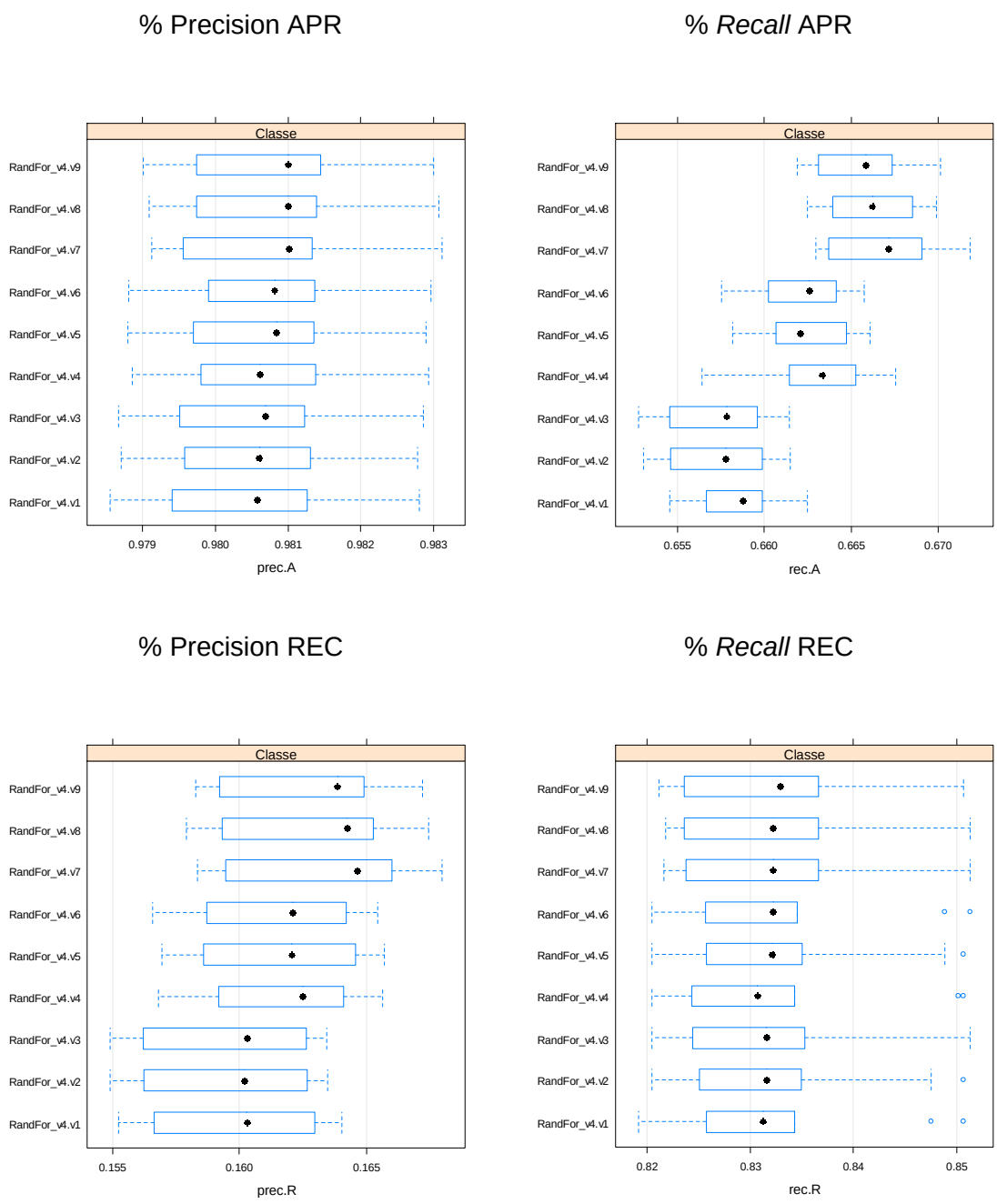


Figura 33 – *Boxplots* das métricas *Precision* e *Recall* para as classes APR e REC, usando *Random Forests*, para a abordagem V4.

4.6. Comparação das abordagens

As diferentes abordagens utilizadas procuram encontrar um modelo ótimo para o processo de decisão Pay No Pay, que se ajuste às características do segmento de

Negócios e que ao mesmo tempo procure sensibilizar os gestores de conta ou cliente para a cultura do risco de crédito, maximizando o lucro.

O Quadro 32 resume os melhores resultados dos modelos que maximizam o lucro, para cada abordagem.

Modelo	*Precision* APR	*Recall* APR	*Precision* REC	*Recall* REC	Erro	Lucro
RandFor_V0.v7	0,9891	0,9982	0,9735	0,8599	0,0118	0,7441
RandFor_V1.v27	0,9986	0,9080	0,4553	0,9840	0,0865	0,8208
RandFor_V2.v3	0,9798	0,6095	0,1438	0,8390	0,3739	0,9954
RandFor_V3.v2	0,9795	0,5983	0,1404	0,8395	0,3842	1,0000
RandFor_V4.v3	0,9805	0,6574	0,1597	0,8331	0,3299	0,9706

Quadro 32 – Comparação de abordagens

Pode-se verificar que, independentemente do método utilizado, a abordagem V0 é aquela que apresenta piores valores para a métrica Lucro, com valores a rondar os 74%. Recorde-se que esta abordagem procura minimizar os erros. A abordagem V1, que testa cortes na probabilidade da classe APR, obtém melhores resultados que a abordagem V0. Porém a abordagem V2, que decide com base previsão ajustada pela matriz de benefícios consegue ainda melhores resultados para a métrica Lucro, do que as abordagens anteriores. A abordagem V3, que consiste na adição de um critério *expert* baseado no descoberto provocado é a abordagem que consegue melhores resultados no que diz respeito à métrica Lucro, com um valor de 100%. A abordagem V4, que utiliza um critério *expert* baseado na relação entre o descoberto provado e o montante da transação, não consegue obter melhor resultado que a abordagem V2.

No que diz respeito à métrica Erro, por observação do Quadro 32, constata-se que a melhor abordagem é a V0, com valores abaixo dos 2%, que é a pior das abordagens para a métrica Lucro. A pior abordagem para a métrica Erro é a abordagem V3, com valores próximos dos 38%.

Os resultados para as métricas *Precision* para a classe APR e REC e *Recall* para a classe APR, acompanham os resultados obtidos para a métrica Erro. Por outras palavras, os melhores resultados são obtidos na abordagem V0, onde a *Precision* para a classe APR e REC ronda os 99% e os 97%, respetivamente, e a métrica *Recall* para a classe APR ronda os 100%. Os piores resultados para estas métricas são obtidos na abordagem V3.

Já a métrica *Recall* para a classe REC, apresenta um comportamento diferente das restantes. O melhor resultado é obtido na abordagem V1 é de 98%, que resulta de um corte de 95% na probabilidade da classe APR.O pior resultado é de 83,31%, que é obtido na abordagem V4. A abordagem V3 tem um resultado semelhante à abordagem V2. O resultado desta métrica é assim influenciado pelo número reduzido de casos na população com a classe REC.

Esta análise permite ainda constatar que nenhum dos modelos é ótimo, ou seja, consegue obter bons resultados em todas as métricas. A maximização da métrica Lucro conduz a maiores valores da métrica Erro e menores valores das métricas *Precision* e *Recall* para cada classe, que não é o pretendido. Neste trabalho foi mais valorizada a métrica Lucro e as abordagens utilizadas procuravam a maximização desta mesma métrica. Com base neste critério, deve ser escolhido o modelo RandFor_V3.v2. Porém, se fosse privilegiada a minimização dos erros e, assim evitar insatisfação do cliente, então o modelo a escolher deve ser o RandFor_V0.v7. Qualquer um dos restantes modelos obtém um resultado intermédio entre estes dois extremos. Dado que cada erro tem custos distintos, pode-se ainda querer tomar uma decisão de forma a evitar erros numa classe em particular. Em particular, se o objetivo fosse identificar corretamente o maior número possível de casos da classe APR deveria ser escolhido o modelo RandFor_V0.v7. Se o objetivo fosse identificar corretamente o maior número possível de casos da classe REC deveria ser escolhido o modelo RandFor_V1.v27.

5. Conclusão

Este trabalho foca a utilização de métodos de classificação para desenvolver um modelo de Pay No Pay (pagar ou não pagar) para o segmento de Negócios.

Foram construídos vários modelos de previsão usando árvores de decisão e *Random Forests*. Estes modelos foram encadeados com modelos de decisão, mantendo-se o objetivo de prever o incumprimento a 22 dias úteis após o pagamento da transação.

Com o objetivo de maximizar o lucro e minimizar os erros cometidos, várias abordagens foram efetuadas e cada tipo de método foi testado usando parâmetros distintos.

O estudo foi iniciado por uma abordagem exclusivamente baseada em modelos de previsão, procurando minimizar o número de erros cometidos, independentemente do impacto que cada erro teria no lucro obtido. Esta abordagem foi denominada de V0. Com o objetivo de testar o impacto de cada tipo de erro em função da confiança associada a cada previsão, a abordagem V1, além do modelo de previsão, inclui um modelo de decisão baseado na probabilidade da classe “aprovação”. Com a abordagem V2, procurou-se decidir a transação com base na confiança associada à previsão, mas tendo em conta a matriz de benefícios para cada transação. Na abordagem V3 é acrescentado um critério *expert* baseado no descoberto provocado pela transação. Na abordagem V4, em semelhança à abordagem V3, é utilizado um critério *expert*, mas neste caso é um critério que relaciona o descoberto provocado com o montante da transação.

Das várias abordagens ao problema resultou que nenhum dos modelos testados obteve resultados ótimos em todas as métricas. A maximização do lucro implicou um aumento dos valores para a métrica Erro, que não era pretendida, e uma diminuição nos valores das métricas *Precision* e *Recall* associadas a cada classe, que também não era pretendida.

Não sendo possível otimizar todas as métricas, privilegiou-se a maximização da métrica Lucro e, para esta métrica, os resultados obtidos permitiram concluir que o método das *Random Forests* é mais estável e obtém melhores resultados do que o método das Árvores de Decisão.

A abordagem V0, baseada apenas no número de erros cometidos, evidenciou que aprovar uma transação quando a mesma deveria ser recusada penaliza mais o lucro do que recusar uma transação que deveria ser aprovada. Por isso, esta abordagem é a que minimiza os erros cometidos, mas não maximiza o lucro.

A abordagem V3, que utiliza um modelo de previsão encadeado com um modelo de decisão que usa uma matriz de benefícios e um critério *expert* baseado no descoberto provocado, é a abordagem que maximiza o lucro, mas não minimiza os erros cometidos.

Os resultados obtidos incentivam a utilização de outras possíveis abordagens ao problema. Uma dessas possíveis abordagens consiste na criação de uma terceira classe, a classe da Revisão Manual, que permitiria selecionar um grupo reduzido de transações, que seriam alvo de decisão manual. A seleção destas transações poderia ser efetuada com base no montante de descoberto provocado e/ou na confiança da previsão efetuada. Uma outra possível abordagem consiste em efetuar a previsão com base em modelos de regressão. Nesta abordagem a previsão teria como resultado o número de dias que o cliente demoraria a regularizar a conta. Outra possível abordagem é a prévia utilização de modelos de *clustering*, para identificar quantas classes poderiam ser criadas com base na regularização de cada transação e aplicação de modelos de previsão e decisão em função das classes previamente identificadas. Outras abordagens possíveis consistem na utilização de outros métodos de previsão, como por exemplo as *Support Vector Machines*.

6. Bibliografia

Breiman, L. (1996a). Bagging predictor. *Machine Learning*, 123-140.

Breiman, L. (1996b). *Out-of-bag estimation.* Obtido em 10 de Janeiro de 2012, de http://ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

Breiman, L. (2001). Random forests. *Machine Learning*, 5-32.

Breiman, L. (1999). *Using adaptive bagging to bias regressions.* Statistics Dept. UCB.

Davies, R. (21 de julho de 2009). Recognition due for business drivers of SEPA. *GTnews*.

Dietterich, T. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 1-22.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern Classification.* Willey - Interscience.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Academic Press INC.

Ghodselahi, A. (Março de 2011). A Hybrid Support Vector Machine Ensemble Model for Credit Scoring. *International Journal of Computer Applications*, pp. 975 - 8887.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence*, 832-844.

Morgan, J. A., & Sonquist, J. N. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association, Vol. 58*, 415-434.

Pliha, R. K. (2004). *Patent No. 858,745.* Nashville TN, US.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 81-106.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. *Machine Learning: An Artificial Inteligence Approach*, 463-482.

Schebesh, K. B., & Stecking, R. (2005). Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 1082-1088.

Sousa, M. R., & Costa, J. P. (2008). A tripartite scorecard for the Pay No Pay decision making in the retail banking industry. *Applications of Data Mining in E-Business and Finance*, 45-50.

Stone, C., Breinam, L., Olshen, R., & Friedman, J. (1984). *Classification and Regression Trees.* Wadsworth Int. Group.

Torgo, L. (2010). *Data mining with R: learning with case studies.* Chapman & Hall/CRC.

7. Anexos

Anexo A – Listagem de características utilizadas

Característica	Descrição
TXN001	Código de transação
TXN002	Montante da transação
TXN003	Código de motivo de rejeição da transação
CTA001	Código de centro de negócio
TXN004	Saldo contabilístico inicial
TXN005	Código de tipo de transação
TXN006	Razão entre o saldo disponível e o montante da transação
TXN007	Razão entre o saldo disponível mais valores em cobrança e o montante da transação
CTA002	Saldo contabilístico inicial
CTA003	Saldo disponível inicial
CTA004	Saldo autorizado inicial
CTA005	Limite autorizado inicial mais valores em cobrança
CTA006	Montante inicial de indisponíveis
CTA007	Limite inicial de descobertos
CTA008	Montante inicial de cativos a débito
CTA009	Montante inicial de cativos a crédito
CTA010	Saldo inicial de crédito automático
CTA011	Saldo inicial da conta investimento
CTA012	Saldo valor inicial
CTA013	Número de dias a descoberto
CTA014	Saldo contabilístico final
CTA015	Saldo disponível final
CTA016	Saldo autorizado final
CTA017	Limite autorizado final mais valores em cobrança
CTA018	Montante final de valores em cobrança
CTA019	Limite final de descobertos
CTA020	Montante final de cativos a débito
CTA021	Montante final de cativos a crédito
CTA022	Saldo final de crédito automático
CTA023	Saldo final da conta investimento
CTA024	Saldo valor final
CTA025	Montante global de transações a decidir
CTA026	Saldo disponível após montante global de transações
CTA027	Status da conta
CTA028	Número de dias no mês em que a conta esteve a descoberto
CTA029	Tipo de conta
CTA030	Indicador de conta ordenado
CTA031	Número de titulares da conta

CTA032	Montante global de valores em cobrança
CTA033	Saldo contabilístico da conta
CTA034	Saldo disponível da conta
CTA035	Saldo valor da conta
CTA036	Valor dos cativos a crédito da conta
CTA037	Valor dos cativos a débito da conta
CTA038	Número de dias de descoberto da conta
CTA039	Número de dias decorridos desde a última transação a crédito
CTA040	Número de dias decorridos desde a última transação a débito
CTA041	Antiguidade da conta
CTA042	Razão entre o saldo disponível e o saldo valor
CTA043	Média do número de dias a descoberto nos últimos 3 meses
CTA044	Número de débitos nos últimos 30 dias
CTA045	Número de débitos nos últimos 60 dias
CTA046	Número de débitos nos últimos 90 dias
CTA047	Montante de débitos nos últimos 30 dias
CTA048	Montante de débitos nos últimos 60 dias
CTA049	Montante de débitos nos últimos 90 dias
CTA050	Número de créditos nos últimos 30 dias
CTA051	Número de créditos nos últimos 60 dias
CTA052	Número de créditos nos últimos 90 dias
CTA053	Montante de créditos nos últimos 30 dias
CTA054	Montante de créditos nos últimos 60 dias
CTA055	Montante de créditos nos últimos 90 dias
CTA056	Número de transações em Pay No Pay na conta no último mês
CTA057	Número de transações em Pay No Pay na conta nos últimos 3 meses
CTA058	Número de transações em Pay No Pay na conta nos últimos 6 meses
CTA059	Número de transações em Pay No Pay na conta nos últimos 12 meses
CTA060	Código de relacionamento cliente-conta
CTA061	Limite de descoberto não utilizado
CTA062	Código de atividade económica
CTA063	Valor da faturação
CTA064	Ano da faturação
CTA065	Segmento do cliente
CTA066	Tipo de cliente
CTA067	Código para identificar segmento do cliente no Pay no Pay
CTA068	Macro-segmento do cliente na atribuição de crédito
CTA069	Score do cliente
CTA070	Limite interno para descoberto
CTA071	Data de validade do limite interno para descoberto
CTA072	Limite potencial de descoberto
CTA073	Limite em contas correntes caucionadas (CCC)
CTA074	Prestações de empréstimos a cobrar no mês

CTA075	Data calendarizada para pagamento da próxima prestação de empréstimos
CTA076	Rating do cliente
CTA077	Contador de incidentes de risco "Muito Graves"
CTA078	Contador de Incidentes de risco "Pouco Graves"
CTA079	Património financeiro médio
CTA080	Rendibilidade do cliente
CTA081	Rendibilidade do cliente nos produtos de passivo
CTA082	Rendibilidade mensal média do cliente
CTA083	Rendibilidade mensal média do cliente nos produtos de passivo
CTA084	Antiguidade do cliente
CTA085	Indicador de bloqueio para mensagens SMS
CTA086	Indicador de bloqueio para mensagens de correio eletrónico
CTA087	Número de transações Pay No Pay do cliente no último mês
CTA088	Número de transações Pay No Pay do cliente nos últimos 3 meses
CTA089	Número de transações Pay No Pay do cliente últimos 6 meses
CTA090	Número de transações Pay No Pay do cliente nos últimos 12 meses
CTA091	Segmento comercial do cliente
CTA092	Código de incidência de risco associada a cheques
CTA093	Número de dias até à expiração dos limites internos
CTA094	Limite disponível na CCC
CTA095	Limite disponível na conta dinâmica
CTA096	Salário do cliente no último mês
CTA097	Salário médio do cliente
CTA098	Saldo disponível final após montante global de transações mais o montante de cativos a débito
CTA099	Razão entre o descoberto provocado e o montante global de valores em cobrança da conta
CTA100	Razão entre o património e o descoberto
CTA101	Razão entre o descoberto e o património
CTA102	Razão entre o último salário do cliente e o descoberto

Quadro 33 – Lista de características

Anexo B – Quadros com os resultados obtidos nas experiências comparativas

Modelo	ntree	Mtry	Est	prec.A	rec.A	prec.R	rec.R	err	% lucro
			Med	0,9885	0,9983	0,9754	0,8517	0,0123	74,06
			DP	0,0005	0,0003	0,0051	0,0059	0,0005	1,55
			Min	0,9876	0,9979	0,9683	0,8436	0,0117	70,59
V0.v1	500	5	Max	0,9894	0,9988	0,9832	0,8607	0,0131	75,98
			Med	0,9885	0,9983	0,9756	0,8518	0,0123	73,82
			DP	0,0005	0,0004	0,0060	0,0057	0,0005	1,48
			Min	0,9876	0,9978	0,9673	0,8436	0,0116	70,75
V0.v2	1000	5	Max	0,9896	0,9989	0,9847	0,8620	0,0132	75,49
			Med	0,9885	0,9984	0,9759	0,8520	0,0123	73,85
			DP	0,0006	0,0004	0,0055	0,0065	0,0006	1,49
			Min	0,9876	0,9979	0,9688	0,8436	0,0115	70,77
V0.v3	1500	5	Max	0,9896	0,9989	0,9847	0,8620	0,0132	75,62
			Med	0,9889	0,9982	0,9738	0,8564	0,0121	74,20
			DP	0,0005	0,0003	0,0043	0,0059	0,0005	1,38
			Min	0,9882	0,9978	0,9676	0,8478	0,0113	71,63
V0.v4	500	7	Max	0,9899	0,9987	0,9811	0,8662	0,0129	75,67
			Med	0,9890	0,9983	0,9747	0,8580	0,0119	74,25
			DP	0,0005	0,0003	0,0043	0,0066	0,0005	1,39
			Min	0,9883	0,9978	0,9676	0,8478	0,0111	71,75
V0.v5	1000	7	Max	0,9901	0,9987	0,9812	0,8682	0,0128	75,66
			Med	0,9890	0,9983	0,9752	0,8574	0,0119	74,20
			DP	0,0005	0,0003	0,0045	0,0066	0,0006	1,37
			Min	0,9881	0,9979	0,9691	0,8464	0,0112	71,68
V0.v6	1500	7	Max	0,9900	0,9988	0,9827	0,8669	0,0130	75,61
			Med	0,9891	0,9982	0,9735	0,8599	0,0118	74,41
			DP	0,0005	0,0003	0,0051	0,0065	0,0005	1,55
			Min	0,9883	0,9977	0,9661	0,8471	0,0111	71,32
V0.v7	500	10	Max	0,9901	0,9988	0,9835	0,8689	0,0129	75,74
			Med	0,9891	0,9982	0,9735	0,8593	0,0119	74,32
			DP	0,0005	0,0003	0,0045	0,0057	0,0005	1,56
			Min	0,9881	0,9978	0,9675	0,8491	0,0114	71,16
V0.v8	1000	10	Max	0,9900	0,9987	0,9819	0,8675	0,0131	75,80
			Med	0,9891	0,9982	0,9742	0,8593	0,0118	74,36
			DP	0,0005	0,0003	0,0044	0,0055	0,0005	1,57
			Min	0,9881	0,9979	0,9685	0,8495	0,0112	71,22
V0.v9	1500	10	Max	0,9900	0,9987	0,9812	0,8669	0,0131	75,78

Quadro 34 – Resultados da abordagem V0 usando *Random Forests*

Modelo	SE	Est	prec.A	rec.A	prec.R	rec.R	Err	% lucro
		Med	0,9842	0,9904	0,8671	0,7967	0,0236	70,59
		DP	0,0010	0,0011	0,0120	0,0138	0,0008	3,35
		Min	0,9826	0,9884	0,8431	0,7741	0,0219	62,08
arvore_v0.v1	0	Max	0,9864	0,9917	0,8813	0,8197	0,0246	72,72
		Med	0,9830	0,9916	0,8797	0,7811	0,0236	69,90
		DP	0,0014	0,0010	0,0125	0,0171	0,0009	3,26
		Min	0,9813	0,9897	0,8569	0,7564	0,0216	61,86
arvore_v0.v2	0.5	Max	0,9857	0,9926	0,8940	0,8107	0,0247	72,54
		Med	0,9816	0,9922	0,8839	0,7623	0,0245	68,67
		DP	0,0010	0,0007	0,0090	0,0124	0,0010	4,51
		Min	0,9797	0,9912	0,8696	0,7426	0,0236	56,90
arvore_v0.v3	1	Max	0,9832	0,9934	0,8968	0,7767	0,0266	72,07

Quadro 35 – Resultados da abordagem V0 usando Árvores de Decisão

Modelo	Ntree	mtry	cut	Est	prec.A	rec.A	prec.R	rec.R	Err	% lucro
				Med	0,9912	0,9965	0,9515	0,8874	0,0114	75,78
				DP	0,0005	0,0004	0,0059	0,0054	0,0005	1,45
				Min	0,9907	0,9959	0,9439	0,8775	0,0104	73,08
V1.v1	500	5	0.6	Max	0,9921	0,9972	0,9617	0,8953	0,0120	77,35
				Med	0,9912	0,9965	0,9517	0,8869	0,0115	75,79
				DP	0,0004	0,0004	0,0057	0,0051	0,0004	1,41
				Min	0,9905	0,9958	0,9423	0,8775	0,0107	73,15
V1.v2	1000	5	0.6	Max	0,9920	0,9972	0,9615	0,8939	0,0122	77,24
				Med	0,9912	0,9965	0,9516	0,8872	0,0115	75,77
				DP	0,0005	0,0004	0,0054	0,0057	0,0004	1,36
				Min	0,9905	0,9958	0,9423	0,8761	0,0108	73,16
V1.v3	1500	5	0.6	Max	0,9920	0,9970	0,9594	0,8946	0,0123	77,11
				Med	0,9915	0,9964	0,9510	0,8910	0,0112	76,04
				DP	0,0005	0,0004	0,0056	0,0063	0,0004	1,36
				Min	0,9905	0,9958	0,9425	0,8799	0,0105	73,60
V1.v4	500	7	0.6	Max	0,9923	0,9971	0,9603	0,8981	0,0120	77,66
				Med	0,9915	0,9964	0,9509	0,8902	0,0113	75,98
				DP	0,0005	0,0004	0,0048	0,0062	0,0004	1,48
				Min	0,9906	0,9959	0,9432	0,8803	0,0106	73,35
V1.v5	1000	7	0.6	Max	0,9923	0,9971	0,9602	0,8988	0,0121	77,52
				Med	0,9915	0,9965	0,9515	0,8905	0,0112	76,01
				DP	0,0005	0,0003	0,0046	0,0061	0,0004	1,40
				Min	0,9906	0,9960	0,9446	0,8803	0,0105	73,34
V1.v6	1500	7	0.6	Max	0,9923	0,9972	0,9610	0,8981	0,0121	77,52
				Med	0,9916	0,9963	0,9493	0,8923	0,0113	76,14
V1.v7	500	10	0.6	DP	0,0004	0,0003	0,0048	0,0053	0,0004	1,34

				Min	0,9911	0,9956	0,9397	0,8824	0,0105	73,83
				Max	0,9922	0,9968	0,9570	0,8982	0,0116	77,54
				Med	0,9916	0,9963	0,9490	0,8916	0,0113	76,11
				DP	0,0004	0,0004	0,0055	0,0052	0,0003	1,47
				Min	0,9910	0,9955	0,9377	0,8824	0,0108	73,46
V1.v8	1000	10	0.6	Max	0,9922	0,9968	0,9558	0,8974	0,0118	77,61
				Med	0,9916	0,9963	0,9490	0,8920	0,0113	76,13
				DP	0,0004	0,0004	0,0053	0,0057	0,0003	1,48
				Min	0,9910	0,9956	0,9390	0,8810	0,0107	73,45
V1.v9	1500	10	0.6	Max	0,9922	0,9968	0,9562	0,8983	0,0117	77,60
				Med	0,9956	0,9841	0,8231	0,9438	0,0188	78,31
				DP	0,0003	0,0011	0,0102	0,0041	0,0009	1,32
				Min	0,9950	0,9823	0,8090	0,9383	0,0178	76,25
V1.v10	500	5	0.8	Max	0,9962	0,9857	0,8371	0,9512	0,0205	79,86
				Med	0,9956	0,9843	0,8244	0,9442	0,0186	78,36
				DP	0,0004	0,0012	0,0104	0,0042	0,0010	1,29
				Min	0,9950	0,9819	0,8067	0,9380	0,0174	76,29
V1.v11	1000	5	0.8	Max	0,9962	0,9858	0,8377	0,9512	0,0209	79,87
				Med	0,9956	0,9842	0,8235	0,9441	0,0187	78,34
				DP	0,0004	0,0011	0,0103	0,0045	0,0009	1,30
				Min	0,9949	0,9822	0,8095	0,9373	0,0175	76,29
V1.v12	1500	5	0.8	Max	0,9962	0,9858	0,8369	0,9512	0,0205	79,89
				Med	0,9956	0,9840	0,8224	0,9444	0,0188	78,37
				DP	0,0004	0,0013	0,0119	0,0041	0,0010	1,34
				Min	0,9949	0,9819	0,8068	0,9366	0,0176	76,44
V1.v13	500	7	0.8	Max	0,9962	0,9856	0,8377	0,9512	0,0208	80,05
				Med	0,9956	0,9842	0,8242	0,9444	0,0186	78,30
				DP	0,0004	0,0012	0,0113	0,0043	0,0010	1,23
				Min	0,9949	0,9820	0,8077	0,9373	0,0174	76,43
V1.v14	1000	7	0.8	Max	0,9963	0,9860	0,8404	0,9525	0,0207	79,85
				Med	0,9956	0,9843	0,8247	0,9443	0,0186	78,31
				DP	0,0004	0,0012	0,0109	0,0046	0,0010	1,25
				Min	0,9949	0,9821	0,8087	0,9373	0,0174	76,37
V1.v15	1500	7	0.8	Max	0,9964	0,9859	0,8396	0,9539	0,0206	79,84
				Med	0,9957	0,9839	0,8210	0,9451	0,0189	78,27
				DP	0,0005	0,0012	0,0109	0,0056	0,0011	1,25
				Min	0,9951	0,9816	0,8038	0,9389	0,0177	76,45
V1.v16	500	10	0.8	Max	0,9965	0,9852	0,8332	0,9559	0,0212	79,94
				Med	0,9956	0,9840	0,8220	0,9448	0,0188	78,27
				DP	0,0004	0,0012	0,0110	0,0053	0,0011	1,22
				Min	0,9949	0,9818	0,8055	0,9373	0,0175	76,48
V1.v17	1000	10	0.8	Max	0,9964	0,9856	0,8369	0,9539	0,0211	79,91
				Med	0,9956	0,9840	0,8221	0,9449	0,0188	78,27
				DP	0,0004	0,0013	0,0118	0,0049	0,0012	1,22
V1.v18	1500	10	0.8	Min	0,9950	0,9816	0,8043	0,9380	0,0174	76,47

				Max	0,9964	0,9858	0,8391	0,9545	0,0212	79,89
V1.v19	500	5	0.95	Med	0,9986	0,9031	0,4426	0,9842	0,0910	81,87
				DP	0,0003	0,0021	0,0077	0,0031	0,0018	0,87
				Min	0,9982	0,9003	0,4313	0,9790	0,0883	80,28
				Max	0,9991	0,9063	0,4537	0,9898	0,0933	82,84
V1.v20	1000	5	0.95	Med	0,9987	0,9029	0,4423	0,9846	0,0911	81,88
				DP	0,0003	0,0021	0,0073	0,0033	0,0018	0,93
				Min	0,9982	0,9001	0,4292	0,9790	0,0890	80,24
				Max	0,9991	0,9056	0,4535	0,9898	0,0935	83,16
V1.v21	1500	5	0.95	Med	0,9987	0,9027	0,4416	0,9844	0,0914	81,88
				DP	0,0003	0,0019	0,0068	0,0034	0,0015	0,93
				Min	0,9981	0,9000	0,4316	0,9783	0,0892	80,23
				Max	0,9991	0,9051	0,4528	0,9898	0,0936	83,18
V1.v22	500	7	0.95	Med	0,9986	0,9060	0,4500	0,9841	0,0884	82,05
				DP	0,0003	0,0023	0,0083	0,0032	0,0021	0,88
				Min	0,9983	0,9027	0,4397	0,9799	0,0852	80,40
				Max	0,9990	0,9097	0,4624	0,9879	0,0911	83,08
V1.v23	1000	7	0.95	Med	0,9986	0,9055	0,4487	0,9843	0,0888	82,02
				DP	0,0002	0,0019	0,0071	0,0028	0,0016	0,90
				Min	0,9983	0,9026	0,4393	0,9803	0,0860	80,26
				Max	0,9991	0,9088	0,4598	0,9891	0,0913	83,16
				Med	0,9986	0,9057	0,4491	0,9841	0,0887	81,91
				DP	0,0002	0,0021	0,0078	0,0026	0,0018	0,90
				Min	0,9984	0,9023	0,4388	0,9810	0,0864	80,26
V1.v24	1500	7	0.95	Max	0,9990	0,9083	0,4602	0,9878	0,0916	83,06
				Med	0,9986	0,9086	0,4569	0,9836	0,0860	82,07
				DP	0,0002	0,0019	0,0077	0,0027	0,0017	0,90
				Min	0,9983	0,9059	0,4477	0,9802	0,0832	80,41
V1.v25	500	10	0.95	Max	0,9990	0,9118	0,4705	0,9878	0,0884	83,18
				Med	0,9986	0,9081	0,4556	0,9840	0,0864	82,08
				DP	0,0002	0,0019	0,0067	0,0025	0,0017	0,84
				Min	0,9983	0,9054	0,4471	0,9802	0,0840	80,60
V1.v26	1000	10	0.95	Max	0,9989	0,9109	0,4665	0,9871	0,0887	83,09
				Med	0,9986	0,9080	0,4553	0,9840	0,0865	82,08
				DP	0,0003	0,0020	0,0068	0,0029	0,0017	0,89
				Min	0,9983	0,9050	0,4457	0,9796	0,0837	80,34
V1.v27	1500	10	0.95	Max	0,9990	0,9113	0,4646	0,9878	0,0891	83,08

Quadro 36 – Resultados da abordagem V1 usando *Random Forests*

Modelo	ntree	mtry	Est	prec.A	rec.A	prec.R	rec.R	err	% lucro
			Med	0,9797	0,6103	0,1439	0,8384	0,3732	99,41
			DP	0,0013	0,0021	0,0026	0,0092	0,0020	0,54
V2.v1	500	5	Min	0,9777	0,6072	0,1399	0,8251	0,3694	98,57

			Max	0,9819	0,6133	0,1479	0,8540	0,3756	100,13
			Med	0,9798	0,6095	0,1438	0,8390	0,3739	99,54
			DP	0,0012	0,0026	0,0026	0,0087	0,0024	0,54
			Min	0,9778	0,6058	0,1397	0,8264	0,3704	98,62
V2.v2	1000	5	Max	0,9818	0,6128	0,1473	0,8540	0,3769	100,24
			Med	0,9798	0,6095	0,1438	0,8390	0,3739	99,54
			DP	0,0013	0,0025	0,0026	0,0093	0,0023	0,55
			Min	0,9778	0,6060	0,1398	0,8264	0,3703	98,62
V2.v3	1500	5	Max	0,9819	0,6126	0,1476	0,8547	0,3767	100,27
			Med	0,9799	0,6149	0,1454	0,8384	0,3689	99,18
			DP	0,0013	0,0027	0,0025	0,0096	0,0024	0,55
			Min	0,9780	0,6093	0,1413	0,8264	0,3649	98,35
V2.v4	500	7	Max	0,9821	0,6181	0,1488	0,8547	0,3725	99,94
			Med	0,9800	0,6144	0,1453	0,8392	0,3693	99,26
			DP	0,0013	0,0019	0,0026	0,0093	0,0018	0,55
			Min	0,9779	0,6116	0,1411	0,8264	0,3659	98,37
V2.v5	1000	7	Max	0,9821	0,6169	0,1495	0,8547	0,3720	100,00
			Med	0,9800	0,6143	0,1453	0,8393	0,3694	99,28
			DP	0,0012	0,0019	0,0026	0,0092	0,0018	0,55
			Min	0,9780	0,6108	0,1412	0,8264	0,3661	98,41
V2.v6	1500	7	Max	0,9821	0,6167	0,1491	0,8554	0,3714	100,03
			Med	0,9801	0,6188	0,1468	0,8393	0,3652	98,97
			DP	0,0013	0,0029	0,0030	0,0099	0,0028	0,57
			Min	0,9783	0,6153	0,1422	0,8277	0,3606	98,05
V2.v7	500	10	Max	0,9823	0,6237	0,1510	0,8554	0,3684	99,79
			Med	0,9801	0,6181	0,1466	0,8394	0,3658	99,06
			DP	0,0013	0,0023	0,0028	0,0094	0,0023	0,53
			Min	0,9783	0,6152	0,1422	0,8277	0,3619	98,24
V2.v8	1000	10	Max	0,9823	0,6217	0,1506	0,8554	0,3691	99,79
			Med	0,9801	0,6179	0,1465	0,8395	0,3661	99,09
			DP	0,0013	0,0023	0,0027	0,0094	0,0022	0,54
			Min	0,9782	0,6148	0,1424	0,8271	0,3617	98,27
V2.v9	1500	10	Max	0,9822	0,6215	0,1504	0,8547	0,3690	99,88

Quadro 37 – Resultados da abordagem V2 usando *Random Forests*

Modelo	SE	Est	prec.A	rec.A	prec.R	rec.R	Err	% lucro
		Med	0,9738	0,6408	0,1452	0,7797	0,3492	95,76
		DP	0,0016	0,0149	0,0064	0,0099	0,0141	1,00
		Min	0,9709	0,6086	0,1318	0,7653	0,3341	93,97
arvore_v2.v1	0	Max	0,9758	0,6570	0,1510	0,7936	0,3801	97,44
		Med	0,9729	0,6229	0,1390	0,7781	0,3658	97,12
		DP	0,0019	0,0133	0,0047	0,0116	0,0125	0,97
		Min	0,9701	0,5958	0,1282	0,7611	0,3547	95,86
arvore_v2.v2	0.5	Max	0,9752	0,6348	0,1437	0,7929	0,3920	98,85
arvore_v2.v3	1	Med	0,9720	0,6049	0,1334	0,7770	0,3826	98,43

		DP	0,0019	0,0178	0,0067	0,0119	0,0167	1,28
		Min	0,9687	0,5778	0,1221	0,7597	0,3570	96,07
		Max	0,9743	0,6319	0,1430	0,7927	0,4084	100,74

Quadro 38 – Resultados da abordagem V2 usando Árvores de Decisão

Modelo	ntree	mtry	Est	prec.A	rec.A	prec.R	rec.R	err	% lucro
			Med	0,9794	0,5988	0,1405	0,8389	0,3838	99,90
			DP	0,0013	0,0021	0,0026	0,0093	0,0020	0,53
			Min	0,9773	0,5963	0,1363	0,8251	0,3803	99,13
V3.v1	500	5	Max	0,9817	0,6015	0,1445	0,8554	0,3864	100,62
			Med	0,9795	0,5983	0,1404	0,8395	0,3842	100,00
			DP	0,0012	0,0026	0,0026	0,0088	0,0024	0,52
			Min	0,9774	0,5948	0,1362	0,8264	0,3811	99,12
V3.v2	1000	5	Max	0,9817	0,6016	0,1440	0,8554	0,3874	100,73
			Med	0,9795	0,5984	0,1404	0,8395	0,3841	100,00
			DP	0,0013	0,0024	0,0026	0,0093	0,0022	0,53
			Min	0,9774	0,5950	0,1363	0,8264	0,3811	99,12
V3.v3	1500	5	Max	0,9817	0,6014	0,1443	0,8561	0,3873	100,75
			Med	0,9796	0,6034	0,1418	0,8389	0,3796	99,68
			DP	0,0013	0,0025	0,0025	0,0096	0,0022	0,53
			Min	0,9776	0,5989	0,1375	0,8264	0,3760	98,89
V3.v4	500	7	Max	0,9819	0,6061	0,1455	0,8561	0,3823	100,47
			Med	0,9796	0,6030	0,1419	0,8397	0,3798	99,74
			DP	0,0013	0,0018	0,0026	0,0094	0,0018	0,54
			Min	0,9775	0,6001	0,1375	0,8264	0,3767	98,90
V3.v5	1000	7	Max	0,9819	0,6052	0,1461	0,8561	0,3827	100,50
			Med	0,9797	0,6031	0,1419	0,8397	0,3797	99,74
			DP	0,0013	0,0018	0,0025	0,0092	0,0017	0,54
			Min	0,9776	0,6004	0,1376	0,8264	0,3767	98,91
V3.v6	1500	7	Max	0,9820	0,6052	0,1458	0,8568	0,3820	100,53
			Med	0,9798	0,6073	0,1432	0,8398	0,3759	99,46
			DP	0,0013	0,0029	0,0029	0,0099	0,0028	0,55
			Min	0,9779	0,6034	0,1385	0,8277	0,3713	98,60
V3.v7	500	10	Max	0,9821	0,6120	0,1472	0,8568	0,3794	100,28
			Med	0,9798	0,6068	0,1430	0,8399	0,3763	99,54
			DP	0,0013	0,0021	0,0027	0,0095	0,0022	0,51
			Min	0,9779	0,6044	0,1386	0,8277	0,3726	98,75
V3.v8	1000	10	Max	0,9821	0,6103	0,1471	0,8568	0,3790	100,29
			Med	0,9798	0,6066	0,1430	0,8399	0,3765	99,57
			DP	0,0013	0,0023	0,0026	0,0094	0,0022	0,52
			Min	0,9778	0,6040	0,1387	0,8271	0,3722	98,79
V3.v9	1500	10	Max	0,9820	0,6100	0,1470	0,8561	0,3789	100,37

Quadro 39 – Resultados da abordagem V3 usando *Random Forests*

Modelo	SE	Est	prec.A	rec.A	prec.R	rec.R	err	% lucro
		Med	0,9735	0,6313	0,1421	0,7802	0,3580	96,09
		DP	0,0016	0,0145	0,0061	0,0099	0,0138	0,99
		Min	0,9706	0,5995	0,1293	0,7659	0,3442	94,34
arvore_v3.v1	0	Max	0,9756	0,6461	0,1478	0,7950	0,3885	97,76
		Med	0,9726	0,6139	0,1362	0,7786	0,3741	97,41
		DP	0,0019	0,0132	0,0045	0,0115	0,0124	1,00
		Min	0,9699	0,5867	0,1258	0,7624	0,3627	96,12
arvore_v3.v2	0.5	Max	0,9750	0,6267	0,1407	0,7943	0,4004	99,18
		Med	0,9716	0,5964	0,1310	0,7774	0,3905	98,69
		DP	0,0018	0,0170	0,0063	0,0118	0,0160	1,25
		Min	0,9684	0,5708	0,1204	0,7611	0,3664	96,34
arvore_v3.v3	1	Max	0,9740	0,6230	0,1397	0,7929	0,4149	100,87

Quadro 40 – Resultados da abordagem V3 usando Árvores de Decisão

Modelo	ntree	mtry	Est	prec.A	rec.A	prec.R	rec.R	err	% lucro
			Med	0,9805	0,6585	0,1600	0,8324	0,3289	96,94
			DP	0,0012	0,0024	0,0032	0,0100	0,0024	0,46
			Min	0,9786	0,6546	0,1553	0,8191	0,3240	96,22
V4.v1	500	5	Max	0,9828	0,6625	0,1640	0,8506	0,3318	97,49
			Med	0,9805	0,6575	0,1597	0,8330	0,3298	97,05
			DP	0,0012	0,0029	0,0032	0,0097	0,0028	0,46
			Min	0,9787	0,6531	0,1549	0,8205	0,3249	96,33
V4.v2	1000	5	Max	0,9828	0,6615	0,1635	0,8506	0,3334	97,56
			Med	0,9805	0,6574	0,1597	0,8331	0,3299	97,06
			DP	0,0012	0,0030	0,0032	0,0100	0,0028	0,47
			Min	0,9787	0,6528	0,1549	0,8205	0,3249	96,32
V4.v3	1500	5	Max	0,9829	0,6615	0,1634	0,8513	0,3332	97,62
			Med	0,9807	0,6629	0,1618	0,8327	0,3248	96,73
			DP	0,0013	0,0032	0,0031	0,0103	0,0029	0,48
			Min	0,9789	0,6564	0,1568	0,8205	0,3193	95,96
V4.v4	500	7	Max	0,9829	0,6676	0,1657	0,8506	0,3292	97,49
			Med	0,9807	0,6622	0,1616	0,8332	0,3254	96,80
			DP	0,0012	0,0026	0,0032	0,0099	0,0026	0,46
			Min	0,9788	0,6582	0,1569	0,8205	0,3206	96,05
V4.v5	1000	7	Max	0,9829	0,6661	0,1657	0,8506	0,3285	97,33
			Med	0,9807	0,6620	0,1616	0,8334	0,3255	96,82
			DP	0,0012	0,0026	0,0031	0,0099	0,0025	0,47
			Min	0,9788	0,6575	0,1566	0,8205	0,3209	96,08
V4.v6	1500	7	Max	0,9830	0,6657	0,1654	0,8513	0,3283	97,38
			Med	0,9808	0,6669	0,1636	0,8333	0,3210	96,50
			DP	0,0013	0,0031	0,0035	0,0105	0,0031	0,48
			Min	0,9791	0,6629	0,1583	0,8216	0,3152	95,71
V4.v7	500	10	Max	0,9831	0,6718	0,1680	0,8513	0,3249	97,02
V4.v8	1000	10	Med	0,9808	0,6660	0,1632	0,8334	0,3219	96,60

			DP	0,0013	0,0027	0,0034	0,0102	0,0027	0,45
			Min	0,9791	0,6625	0,1579	0,8218	0,3171	95,85
			Max	0,9831	0,6699	0,1675	0,8513	0,3259	97,11
			Med	0,9808	0,6656	0,1630	0,8334	0,3222	96,63
			DP	0,0013	0,0027	0,0033	0,0101	0,0027	0,46
			Min	0,9790	0,6619	0,1583	0,8211	0,3169	95,86
V4.v9	1500	10	Max	0,9830	0,6701	0,1672	0,8506	0,3261	97,12

Quadro 41 – Resultados da abordagem V4 usando *Random Forests*

Modelo	SE	Est	prec.A	rec.A	prec.R	rec.R	Err	% lucro
		Med	0,9750	0,6865	0,1622	0,7749	0,3071	93,46
		DP	0,0015	0,0148	0,0079	0,0108	0,0141	0,98
		Min	0,9723	0,6554	0,1462	0,7598	0,2920	91,93
arvore_v4.v1	0	Max	0,9772	0,7026	0,1691	0,7916	0,3370	95,25
		Med	0,9742	0,6694	0,1547	0,7732	0,3230	94,76
		DP	0,0018	0,0129	0,0057	0,0125	0,0123	0,99
		Min	0,9716	0,6434	0,1420	0,7558	0,3105	93,45
arvore_v4.v2	0.5	Max	0,9767	0,6827	0,1610	0,7909	0,3482	96,61
		Med	0,9734	0,6522	0,1481	0,7719	0,3391	95,99
		DP	0,0018	0,0178	0,0081	0,0129	0,0168	1,29
		Min	0,9703	0,6236	0,1341	0,7538	0,3128	93,45
arvore_v4.v3	1	Max	0,9756	0,6801	0,1601	0,7902	0,3663	98,07

Quadro 42 – Resultados da abordagem V4 usando Árvores de Decisão