Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Predição da recuperação da inadimplência em operações de crédito

Rogério Gomes Lopes

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Brasília
2017

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

GR723p

Gomes Lopes, Rogério
Predição da recuperação da inadimplência em operações de crédito / Rogério Gomes Lopes; orientador Marcelo Ladeira. - Brasília, 2017.
114 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2017.

1. Mineração de Dados. 2. Recuperação de Crédito. 3. Aprendizagem de Máquina. 4. Classificação. 5. Modelo Preditivo. I. Ladeira, Marcelo, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Predição da recuperação da inadimplência em operações de crédito

Rogério Gomes Lopes

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Thiago de Paulo Faleiros
CIC/UnB

Prof. Dr. Remis Balaniuk
UCB

Prof. Dr. Guilherme Novaes Ramos
CIC/UnB

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 31 de agosto de 2017

Dedicatória

Estas foram algumas das mensagens que recebi quando, no dia do meu aniversário em 2015, avisei à minha família que estava concorrendo a uma vaga para o mestrado e, posteriormente, quando submeti um artigo para a Conferência em Los Angeles:

"Vai ser um presente de aniversário."

"Estaremos pedindo a Deus que conceda o seu desejo."

"Se Deus quiser essa vaga será sua, amor. Vamos orar."

"Vai dar certo!"

"Que bom, filho! Você gosta de desafios!"

"Com a ajuda de Deus você vai vencer essa batalha."

"E vencerá com louvor!!!"

Esse trabalho é dedicado a vocês, que sempre confiaram em mim, apoiaram-me sem restrições e foram minha fonte de inspiração para todo esforço. Meu pai Dionísio, minha mãe Nilva, minha esposa Luciana, minha filha Isabella, minha irmã Raquel, minha irmã Vanessa, meu cunhado Alex e meu quase cunhado Rodrigo.

Agradecimentos

Foram longos 755 dias desde o primeiro dia de aula no módulo 14 até o dia da defesa desta dissertação no LARA. Ainda que haja 95% de confiança de que me esquecerei de citar alguns dos meus ajudadores, não posso fugir dessa grandiosa tarefa.

Começo agradecendo o apoio que recebi da instituição onde trabalho, na figura dos meus gerentes, que possibilitaram minha dedicação a esta pesquisa. André, Fabiana, Marcelo, Sandro, Auro, Geni e Luciana.

Em uma conversa que não durou mais do que 5 minutos, um colega me fez um convite para participar de um grupo de trabalho com a missão de sugerir soluções para o problema da PCLD, que foi a grande inspiração para este trabalho. Ao Toncas, deixo meu agradecimento.

A todos os colegas de trabalho e amigos que me apoiaram, incentivaram, comemoraram a cada pequena conquista obtida nesta pesquisa e foram compreensivos com minhas ausências.

Aos colegas do mestrado, um grande ativo conquistado, formamos um grupo espetacular. Foi uma das melhores partes do nosso curso. Aloísio, Caio, Ebberth, Eduardo, Lucas, Sílvio e 4D.

Participar de uma conferência internacional na Califórnia foi uma grande realização e só foi possível pelo apoio recebido dos meus pais Dionísio e Nilva, e o colega Sílvio. Um agradecimento especial pela generosidade de vocês.

Aos professores do PPCA, que muito nos inspiraram a enfrentar esse desafio. De modo especial, agradeço ao meu orientador Professor Marcelo Ladeira, que me auxiliou a seguir os difíceis caminhos deste estudo, cobrou-me muito empenho, mas demonstrou grande compreensão quando foi preciso.

À minha família, em especial minha esposa Luciana, minha filha Isabella, meus pais Dionísio e Nilva, minhas irmãs Raquel e Vanessa. Vocês me apoiaram em todos os momentos nesses 2 anos, torceram por mim, incentivaram e foram compreensíveis com minhas ausências. Só consegui me concentrar neste estudo porque vocês foram meu suporte.

A Deus, por dar-me saúde e me cercar dessa grande quantidade de colegas, amigos e familiares.

Resumo

Este trabalho propôs a indução de classificadores, a partir da aplicação de técnicas de mineração de dados, para identificar clientes inadimplentes com potencial de regularização da dívida visando auxiliar uma instituição financeira a reduzir a Provisão para Créditos de Liquidação Duvidosa (PCLD). Estes modelos poderão contribuir para reversão de despesas da instituição financeira. Foram utilizados as técnicas Generalized Linear Models (GLM), Distributed Random Forest (DRF), Deep Learning (DL) e Gradient Boosting Methods (GBM), implementados na plataforma H2O.ai. Alguns aspectos que afetam o comportamento do cliente inadimplente foram identificados, como o perfil de sua renda e a época do ano. Estratégias de recuperação de crédito foram propostas e simulações identificaram possibilidades de redução de despesas operacionais.

Palavras-chave: Mineração de Dados, Inadimplência, PCLD, Aprendizagem de Máquina, GBM, DRF, Boosting, Random Forest, Deep Learning.

Abstract

This works proposes the induction of classifiers, from the application of data mining techniques, to identify defaulting clients with debt settlement potential to assist a financial institution in reducing its provision for doubtful debits. These models may contribute to the reversal of expenses of the financial institution. The techniques Generalized Linear Models (GLM), Distributed Random Forest (DRF), Deep Learning (DL) and Gradient Boosting Methods (GBM) algorithms implemented in the H2O.ai platform were used. Some aspects that affect the behavior of the defaulting customer, such as the profile of their income and the period of the year, have been identified. Strategies of credit recovery strategies were proposed and simulations identified possibilities of reducing operating expenses.

Keywords: Data Mining, Loan Default, Machine Learning, GBM, DRF, Boosting, Random Forest, Deep Learning.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ASTESJ *Advances in Science, Technology and Engineering Systems Journal.*

BACEN Banco Central do Brasil.

COPOM Comitê de Política Monetária do BACEN.

CRISP-DM *Cross Industry Standard Process for Data Mining.*

CRM Gestão do Relacionamento com o Cliente.

DL *Deep Learning.*

DRF *Distributed Random Forest.*

GBM *Gradient Boosting Machine.*

GLM *Generalized Linear Models.*

ICMLA IEEE - *Internacional Conference on Machine Learning and Applications.*

LASSO *Least Absolute Shrinkage and Selection Operator.*

PCLD Provisão para Crédito de Liquidação Duvidosa.

PF Pessoa Física.

SFN Sistema Financeiro Nacional.

SGS Sistema Gerenciador de Séries Temporais.

Capítulo 1

Introdução

Este capítulo apresenta o problema proposto como tema da pesquisa de mestrado, abordando sua definição e justificativa. Em seguida, são detalhados os objetivos, hipóteses e contribuições esperadas.

1.1 Definição do Problema

Este estudo propõe-se a elaborar um modelo preditivo para auxiliar na identificação dos clientes com maior potencial de regularização de suas operações de crédito em atraso em uma instituição financeira brasileira.

1.1.1 O Aumento da Inadimplência

Desde janeiro de 2015, observou-se uma queda na oferta de crédito e o aumento da inadimplência em operações de crédito, no âmbito do Sistema Financeiro Nacional (SFN), resultante da diminuição da atividade econômica e confiança dos investidores. Segundo o Banco Central do Brasil (BACEN), a inadimplência das operações de Pessoa Física (PF) apresentou um crescimento de 13,5%, saindo de 3,7% em dezembro de 2014 e alcançando 4,2% em setembro de 2016, representando um aumento de 0,5 pontos percentuais no período.[1]

A Figura 1.1 apresenta a série mensal da inadimplência observada entre os clientes PF entre os meses de dezembro de 2014 a setembro de 2016. Além do crescimento, é possível observar que a inadimplência apresentou alguns meses de recuperação, isto é, uma diminuição da inadimplência, possivelmente causada pela sazonalidade de recebimentos de proventos adicionais no final do ano, como o 13º salário ou férias.

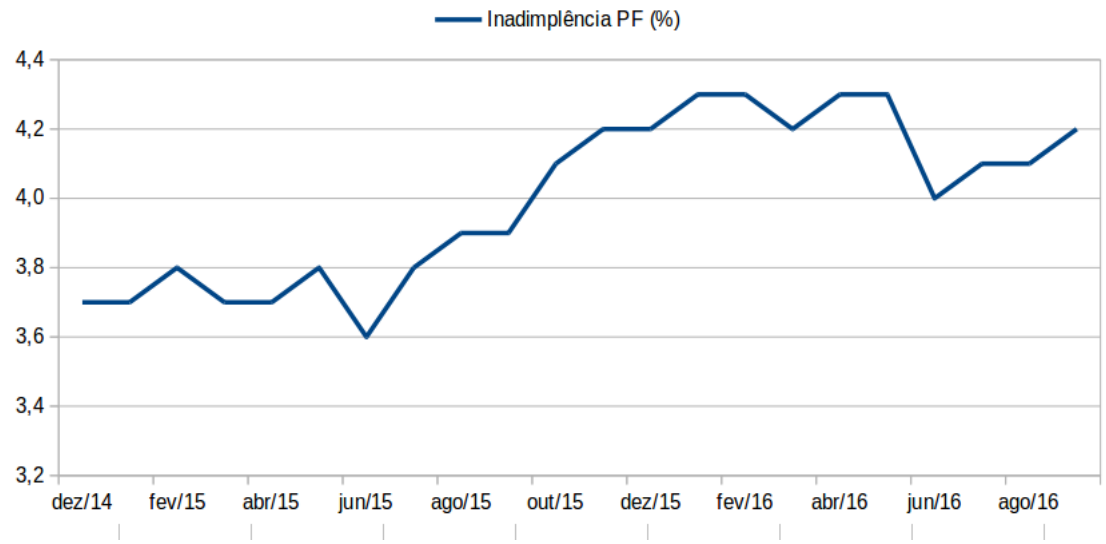

Figura 1.1: Aumento da Inadimplência de Pessoas Física.

No âmbito da instituição financeira estudada neste trabalho, que será identificada apenas como Banco Alfa, apesar de apresentar inadimplência historicamente menor do que o SFN, também observou-se um aumento dos atrasos nos pagamentos das operações de crédito.

O Banco Alfa possuía aproximadamente 54 milhões de contratos de operações de crédito ativas com pessoas físicas no final de julho de 2016. Deste montante, cerca de 8.6 milhões estavam com atraso igual ou superior a 14 dias, representando 15,9% dos contratos. Estes contratos em atraso atingiram mais de R$ 20,8 bilhões, representando cerca de 5,8% da carteira de operações de crédito de pessoas físicas desse banco, um aumento de 1,2 pontos percentuais em relação ao observado em dezembro de 2014. Ou seja, em 21 meses o volume financeiro de operações de crédito em atraso contratadas por pessoas físicas aumentou 23%.

A Resolução 2.682 de 22/12/1999 do BACEN determina que as instituições financeiras classifiquem suas operações de crédito e realizem uma Provisão para Crédito de Liquidação Duvidosa (PCLD), de acordo com uma classificação de risco. Dentre os critérios estabelecidos por esta resolução, está a quantidade de dias em atraso, que estabelece o nível mínimo de risco no qual a operação deverá ser classificada. A Tabela 1.1 apresenta as faixas de dias em atraso consideradas para determinar a classificação de risco e, consequentemente, o percentual mínimo de PCLD que deve ser realizado pelas instituições financeiras. À medida que uma operação aumenta em quantidade de dias em atraso, há um aumento da PCLD, não-linear, que pode atingir 100% do saldo devedor do contrato. Por exemplo, uma operação com um saldo devedor de R$ 1.000, com 15 dias em atraso, deverá realizar um provisionamento mínimo de R$ 10. O valor da provisão poderá atingir

o valor total do saldo devedor, caso seja atingido 180 dias de atraso.

Tabela 1.1: Dias de atraso x Risco x % de Provisão

Dias em atraso	Classificação Mínima	% Provisão
15 a 30	B	1
31 a 60	C	3
61 a 90	D	10
91 a 120	E	30
121 a 150	F	50
151 a 180	G	70
acima de 180	H	100

Após realizar a classificação de risco de cada operação de um cliente, ainda é necessário reclassificar todas as suas operações conforme a operação de pior risco. Esta reclassificação é conhecida como Efeito de Arrasto, quando uma operação recebe uma classificação de risco pior, em decorrência de outra operação. A Tabela 1.2 ilustra um exemplo deste Arrasto, onde uma operação de menor valor, mas com maior atraso, provoca uma piora no risco de outras operações do mesmo cliente. Neste exemplo, o atraso de 62 dias de operação C, não apenas provoca uma PCLD de 10% em sua operação, mas também a aplicação desta alíquota em todas as demais.

1.1.2 Manutenção da Carteira de Clientes

No momento da concessão de crédito, as instituições financeiras assumem o risco de crédito e realizam os respectivos provisionamentos conforme a legislação vigente. Agindo desta maneira, em uma eventual inadimplência do cliente, a instituição financeira estará protegida, assim como a estabilidade do sistema financeiro. Porém, à medida que um cliente atrasa suas operações, a reação natural das instituições financeiras é restringir o crédito a eles, aumentando as chances de evasão destes clientes para outras instituições, uma vez que eles não conseguirão realizar novas operações de crédito com esta instituição.

Com o aumento da inadimplência, visando mitigar a evasão de seus clientes, iniciou-se uma mobilização dos gerentes de conta do banco, com o objetivo de abordar os clientes que apresentavam atrasos em suas operações, propondo alternativas que pudessem sanar

Tabela 1.2: Exemplo do Efeito Arrasto

Operação	Atraso em Dias	Valor	Risco Original	PCLD Original	Risco Contabilizado	PCLD Contabilizada
A	15	80.000	B	800	D	8.000
B	32	10.000	C	300	D	1.000
C	62	2.500	D	250	D	250

o atraso de suas operações, resolvendo a situação de inadimplência e a possível perda do cliente de sua carteira, além de reduzir o montante financeiro destinado à PCLD.

Para cumprir esta tarefa, a instituição disponibilizou um painel de acompanhamento da PCLD para os gerentes de contas, com a listagem de todos os seus clientes com operações em atraso, em ordem decrescente de despesa de provisão, isto é, os primeiros clientes a serem abordados eram aqueles com a maior despesa prevista. Consequentemente, de maneira geral, estes também eram os clientes com as operações com o maior tempo em atraso, em função dos critérios estabelecidos pela Resolução 2.682.

Apesar de não existir um registro oficial do acompanhamento destas ações, houve vários relatos de dificuldades enfrentadas pelos gerentes de conta na execução desta estratégia. Em linhas gerais, enfrentou-se duas grandes dificuldades: a incapacidade financeira do cliente para regularizar as operações em atraso e a impossibilidade do gerente de contas abordar todos os seus clientes listados no painel de acompanhamento da PCLD.

Identificação dos Clientes com Potencial de Redução no Atraso

Diante desta situação, o cenário ideal seria a disponibilização da lista de clientes devedores ordenada pelo potencial de recuperação estimado para o cliente, apoiado em um modelo preditivo.

Apesar de existirem vários estudos para identificar o risco de crédito de um cliente, qualificando-os como bons ou maus pagadores, auxiliando na tomada de decisão para a concessão de crédito, uma vez ocorrida a inadimplência, há poucas pesquisas estudando o risco deste cliente vir a se tornar adimplente novamente.[2]

Lessmann *et al.*[3] publicaram um estudo em outubro de 2015, realizando uma revisão do estado da arte da classificação do risco de crédito. Neste estudo constataram que, apesar de haver muitas pesquisas para elaboração de modelos de concessão de crédito, poucas têm explorado os recentes avanços obtidos no aprendizado preditivo, como a utilização de aprendizagem de máquina. Além disto, muitos estudos são feitos com bases de dados com poucas variáveis e poucos exemplos para treinamento, validação e teste. Os algoritmos usualmente aplicados nesta temática são regressão logística e árvores de decisão, com suas variantes de utilização de *boosting*, *bagging* e florestas.

Porém, há outras técnicas utilizadas para tarefas de classificação, que vêm apresentando excelentes resultados como a *Deep Learning*. LeCun *et al.*[4] apontaram que esta técnica ainda pode apresentar muitos avanços em outros domínios do conhecimento, especialmente em problemas de alta dimensionalidade. O problema da recuperação de crédito pode estar associado a inúmeros motivos, não apenas características demográficas dos clientes, mas também comportamentais ou macroeconômicos, caracterizando-se como um problema candidato a ser bem resolvido com o uso da técnica *Deep Learning*.

1.2 Justificativa do Tema

O tema deste estudo justifica-se por diversas perspectivas, listadas a seguir:

Reversão de Despesa de PCLD - Este estudo pode contribuir para a redução da atual despesa com PCLD, representando um ganho tangível e, ainda, com a manutenção do cliente na carteira da instituição, que pode ser considerado um benefício intangível. Esta redução, além de representar um ganho financeiro, poderá contribuir com a valorização dos ativos do Banco Alfa na Bolsa de Valores.

Especificidade dos Modelos e Interesses Comerciais - A elaboração de modelos preditivos para recuperação de crédito possui características muito específicas em cada instituição e, associado ao interesse comercial do tema, percebe-se poucas publicações a respeito deste assunto.

1.3 Hipóteses de Pesquisa

Algumas hipóteses foram formuladas, que direcionarão a realização deste estudo, e estão listadas a seguir:

- **O perfil de renda do cliente afeta a inadimplência.** Estima-se que o potencial de recuperação de um cliente esteja associado ao segmento de renda a que ele pertença.
- **O produto de crédito afeta a inadimplência** - Para um mesmo cliente, acredita-se que o potencial de recuperação da inadimplência seja diferente em razão do produto consumido. Por exemplo, é possível que financiamentos imobiliários tenham maior prioridade para regularização do que um crédito direto ao consumidor, em virtude das consequências da inadimplência em cada um deles.
- **O modelo preditivo para recuperação de crédito é afetado pelo período do ano.** Em razão da natureza sazonal das receitas e despesas de clientes pessoas físicas, como recebimento de férias, 13º salário, pagamento de impostos, matrículas escolares, supõe-se que o modelo preditivo tenha forte influência do mês em que está ocorrendo a análise do potencial de recuperação do cliente.

1.4 Objetivo Geral

Este estudo visa apoiar o Banco Alfa na redução da PCLD, auxiliando na identificação dos clientes inadimplentes com o maior potencial de regularização de suas operações.

1.5 Objetivos Específicos

Para atingir o objetivo geral deste trabalho, foram definidos os seguintes objetivos específicos:

1. Elaborar modelos preditivos de acordo com o perfil de renda do cliente e compará-los a um modelo preditivo que não considere o perfil de renda do cliente.

2. Elaborar modelos preditivos conforme o tipo de produto consumido e compará-los a um modelo preditivo único, independente do produto.

3. Elaborar modelos preditivos para cada mês de observação e compará-los a um modelo preditivo que leve em conta as observações independente do mês.

4. Criar um índice para auxiliar na prioridade de abordagem para tentativa de renegociação das operações de crédito em atraso.

5. Estimar os montantes financeiros que podem ser recuperados em operações de crédito de pessoas físicas que se encontram em situação de inadimplência.

1.6 Estrutura deste Documento

O Capítulo 2 resume o estado da arte utilizado como direcionador deste trabalho. O Capítulo 3 detalha metodologia aplicada para alcançar os objetivos propostos. O Capítulo 4 evidencia as etapas realizadas no entendimento e preparação dos dados, detalhando os resultados das análises descritivas e as bases de dados utilizadas nesta pesquisa. O Capítulo 5 exibe os procedimentos realizados para a modelagem e as validações dos modelos produzidos. O Capítulo 6 expõe as análises dos resultados obtidos por esse trabalho e discute as possibilidades de uso dos modelos preditivos elaborados. O Capítulo 7 apresenta as conclusões, resultados obtidos e trabalhos futuros. Os Apêndices A e B trazem as cópias dos artigos publicados no IEEE - *Internacional Conference on Machine Learning and Applications* (ICMLA) e *Advances in Science, Technology and Engineering Systems Journal* (ASTESJ).

Capítulo 2

Revisão do Estado da Arte

Este capítulo apresenta o estado da arte da concessão de crédito, que foi utilizado como referência para este estudo, em razão da semelhança com o tema de recuperação de crédito, visto que na pesquisa realizada, não foram encontradas publicações específicas sobre o tema dessa pesquisa.

A busca por trabalhos acadêmicos que abordavam o problema da recuperação de crédito com técnicas de mineração de dados resultou na identificação de um único autor [2]. Nesse trabalho, a predição da recuperação de crédito limitava-se aos contratos de cartão de crédito, enquanto o cenário do Banco Alfa envolvia uma grande quantidade de tipos de operações de crédito, onde cada um deles possui regras distintas de negociação da inadimplência.

Contudo, há uma grande semelhança entre a predição da recuperação de crédito e a análise da concessão de crédito. Em comum, têm-se a necessidade de classificar clientes quanto à sua capacidade de cumprir compromissos de pagamentos de contratos de crédito. Outra semelhança está a alta dimensionalidade do problema, onde identifica-se muitas características que influenciam nesse comportamento do cliente. Entretanto, invertem-se suas classes majoritárias e minoritárias. Enquanto na concessão de crédito a classe majoritária é a de bons pagadores, na recuperação de crédito esta classe torna-se minoritária, ou seja, aqueles que conseguirão sair da inadimplência e retornar à condição de bons pagadores.

Em razão das semelhanças encontradas e na escassez de trabalhos acerca da recuperação de crédito, utilizou-se o estado da arte da análise de concessão de crédito para direcionar esta pesquisa.

Os modelos de análise da concessão de crédito são desenvolvidos para estimar a probabilidade de um cliente apresentar um comportamento indesejado no futuro, como o não pagamento de suas obrigações. Em 2015, Lessmann *et al.* [3] publicaram um artigo com a comparação de diversos algoritmos utilizados em classificação do risco de crédito [5]. Neste

Tabela 2.1: Bases de Dados Utilizadas na Comparação do Estado da Arte.

Nome	**Tamanho da Amostra**	**Variáveis Independentes**	**Taxa de Inadimplência**
AC	690	14	0.445
GC	1000	20	0.300
Th02	1225	17	0.264
Bene 1	3123	27	0.667
Bene 2	7190	28	0.300
UK	30000	14	0.040
PAK	50000	37	0.261
GMC	150000	12	0.067

estudo, foram avaliados 41 algoritmos de classificação utilizados em trabalhos que foram publicados entre 2006 e 2014, agrupados em 3 categorias: i) classificadores individuais, ii) multiclassificadores homogêneos e iii) multiclassificadores heterogêneos. A categoria i) era composta por classificadores que utilizam um único modelo de preditivo. Já as categorias ii) e iii) eram compostas de classificadores que utilizavam uma combinação (*ensemble*) de modelos. Nos homogêneos, a composição era realizada com modelos induzidos com um único algoritmo, enquanto os heterogêneos eram compostos por modelos induzidos com mais de um algoritmo.

A Tabela 2.1 lista as oito bases de dados que foram utilizadas no estudo, em cada um dos 41 algoritmos, cujos modelos foram avaliados sob a perspectiva de 6 indicadores: área sob a curva ROC (AUC), percentual de classificações corretas (Acurácia), índice de Gini parcial (Gini), *H-measure*, *Brier Score* (BS) e Kolmogorov-Smirnov (KS).

A AUC permite avaliar um modelo comparando a proporção de falsos positivos à medida em que é aumentada a taxa de verdadeiros positivos.[6]. A Acurácia é uma medida de precisão geral, indicando a percentagem total de classificações corretas .[7] [3].

Os algoritmos foram comparados segundo sua performance diante de cada um dos 6 indicadores, ao longo de cada uma das 8 bases de dados. Uma nota foi atribuída a cada algoritmo, referente à classificação recebida na comparação entre eles. Por exemplo, o algoritmo *K-means* ficou em 10º lugar considerando o indicador AUC, enquanto o KNN ficou em 34º lugar. Assim, as notas atribuídas a eles foram 10 e 34, respectivamente. Ao final, os algoritmos foram ordenados pela média das notas obtidas, ficando em 1º lugar o algoritmo que obteve as menores notas.

Ao final do estudo, os algoritmos que utilizaram multiclassificadores heterogêneos apresentaram um melhor desempenho global, ainda que a diferença de desempenho entre as três categorias tenha sido pequena. Considerando a simplicidade dos classificadores individuais, eles poderiam ser usados, fornecendo resultados semelhantes aos algoritmos mais complexos.

A Tabela detalhada apresenta os resultados do *benchmark*, trazendo novas referências de desempenho e reconhecimento de novos algoritmos. Como pode ser visto nesta Tabela, o modelo HCES-Bag[8] obteve o maior resultado AUC, enquanto os modelos AVGW e Gasen alcançaram 80,7 % de Acurácia.

Tabela 2.2: Estado da Arte - Comparação dos Modelos

	Algoritmo	**AUC**	**Acurácia**
Multiclassificador Heterogêneo	HCES-Bag	0.932	80.2
	AVG W	0.931	80.7
	GASEN	0.931	80.7
Multiclassificador Homogêneo	RF	0.931	78.9
	BagNN	0.927	80.2
	Boost	0.930	77.2
Individual	LR	0.931	70.8
	LDA	0.929	78.4
	SVM-Rbf	0.925	79.9

Capítulo 3

Metodologia de Pesquisa

Este capítulo apresenta a metodologia proposta para o presente trabalho, voltada para atingir os objetivos declarados. A pesquisa foi organizada em fases e, quando aplicável, foram organizadas conforme as fases propostas pelo *Cross Industry Standard Process for Data Mining* (CRISP-DM) [9].

Nas próximas seções serão detalhadas as fases deste trabalho, a saber: Entendimento do negócio, Compreensão dos dados, Preparação de dados, Modelagem, Avaliação e Implementação.

Estruturação das bases de dados - Esta etapa teve a finalidade de obter, compreender, preparar e disponibilizar as bases de dados para a realização dos treinamentos, validações e testes dos modelos preditivos. Uma vez que o aumento da inadimplência foi observado a partir de janeiro de 2015.[1], foram coletados dados desde esta data até dezembro de 2016, perfazendo um total de 24 meses de estudo, cobrindo as eventuais particularidades no comportamento dos clientes ao longo de um ano.

1. **Obtenção dos dados** - Nesta atividade, foram buscadas três fontes de dados distintas para serem utilizadas na construção dos modelos preditivos, a saber:

    1.1. base de dados das operações de crédito em atraso de pessoas físicas.

    1.2. base de dados cadastrais e comportamentais dos clientes existentes à época da observação dos atrasos.

    1.3. base de dados de indicadores econômicos.

2. **Entendimento dos dados** - O objetivo desta atividade era a compreensão dos dados por meio da realização de uma análise descritiva, com a aferição da qualidade dos dados, identificação da existência de dados faltantes, valores fora da curva (*outliers*) e seleção de variáveis relevantes para o modelo.

3. **Preparação dos dados** - Nesta atividade foi realizada a limpeza das bases de dados, retirada de eventuais dados inconsistentes e elaboração de novas bases de dados, derivadas da base original, segregadas por produtos e perfil de renda dos clientes, a saber:

As tarefas de entendimento e preparação de dados foram realizadas por meio da linguagem R, versão 3.3.2, instalada em um servidor com 34 núcleos de processamento, 64 GB de memória RAM e sistema operacional Redhat 6.

Modelagem e avaliação dos modelos - Nesta etapa foram realizadas as fases de modelagem e avaliação dos modelos:

1. **Modelagem** - A modelagem foi realizada em dois grandes blocos: 1) Modelagem Mensal - treinamento com informações limitadas de um mês específico de 2015 e validação com o respectivo mês em 2016 e 2) Modelagem Anual - treinamento com os 12 (doze) meses de 2015 e validado com cada um dos meses de 2016.

2. **Avaliação dos modelos** - Apuração das métricas de performance de modelos nas bases de dados.

Os modelos foram induzidos na plataforma de aprendizagem de máquina H2O.ai [1], por meio de sua interface com o R. Foram utilizadas 4 máquinas para a formação de um *cluster*, totalizando 170 núcleos de processamento e 200 GB de memória RAM.

Análise dos resultados obtidos - Nesta etapa, foram confrontados os resultados obtidos nas etapas anteriores, buscando validar os modelos comparando as estratégias de segmentação das bases por produto, perfil da renda do cliente, treinamento anual e mensal. Também fez parte desta análise a comparação dos procedimentos utilizados atualmente pelo Banco Alfa e o melhor modelo obtido neste estudo.

Implementação - Esta etapa será dedicada para detalhar as atividades necessárias para a implementação do modelo no processo operacional do Banco Alfa. Será utilizada a plataforma H2O.ai para a integração dos modelos vencedores com o processo operacional do banco. Esta integração consiste no acréscimo ao painel de monitoramento de PCLD, ferramenta já disponível e utilizada pelo Banco Alfa, de um índice que representa a chance de renegociação da dívida em atraso do cliente. Todavia, a sua execução depende dos resultados obtidos na etapa anterior, com a comprovação da superioridade do modelo desenvolvido.

[1] h2o.ai - `http://www.h20.ai`

Capítulo 4

Entendimento e Preparação dos Dados

Este capítulo apresenta as atividades realizadas para a obtenção, entendimento e preparação dos dados. Nas seções seguintes serão apresentados os resultados destas tarefas e algumas evidências de suas realizações.

4.1 Obtenção das Bases de Dados

As bases de dados foram obtidas por meio da extração de dados de sistemas legados da instituição, de *Data Marts* oriundos da área de Gestão do Relacionamento com o Cliente (CRM) e de dados macroeconômicos publicados pelo BACEN, abrangendo o período de janeiro de 2015 a dezembro de 2016.

As bases das operações de crédito em atraso, foram extraídas do sistema de cálculo de PCLD da instituição e representa uma visão contábil das operações, tendo os seus valores auditados pelo órgão regulador BACEN, contendo apenas as informações relativas ao contrato de crédito, como o produto contratado, quantidade de dias em atraso, classificação de risco, valor de despesa de PCLD, entre outros.

Os dados cadastrais e comportamentais dos clientes foram identificados em uma única fonte, o *data mart* corporativo da instituição. Trata-se de um *data mart* desenvolvido para atender um projeto de CRM, com variáveis referentes a dados demográficos como idade, gênero, estado civil, nível de instrução, renda salarial, município de residência, valor da margem de contribuição e quantidade de filhos. Este *data mart* possui um histórico mensal de informações, o que nos permitirá associar os dados dos contratos com os dados cadastrais do cliente à época da identificação dos atrasos de suas operações.

Tabela 4.1: Tabela de Indicadores Econômicos (%)

Mês	Selic (%) (4189)	Ipca (%) (433)	Desocupação (%) (24.369)	Taxa Média de Juros(%) (20.716)	Cesta Básica (R$) (206)
jan/2015	11,82	1,24	6,8	31,97	421,51
fev/2015	12,15	1,22	7,4	32,84	424,45
mar/2015	12,58	1,32	7,9	33,13	425,14
abr/2015	12,68	0,71	8,0	33,88	429,28
mai/2015	13,15	0,74	8,1	34,82	437,11
jun/2015	13,58	0,79	8,3	35,54	442,66
jul/2015	13,69	0,62	8,6	36,49	437,92
ago/2015	14,15	0,22	8,7	37,08	438,30
set/2015	14,15	0,54	8,9	37,57	445,80
out/2015	14,15	0,82	8,9	38,67	451,12
nov/2015	14,15	1,01	9,0	39,02	469,62
dez/2015	14,15	0,96	9,0	38,04	487,27
jan/2016	14,15	1,27	9,5	39,58	646,09
fev/2016	14,15	0,90	10,2	40,23	649,72
mar/2016	14,15	0,43	10,9	40,95	660,38
abr/2016	14,15	0,61	11,2	41,82	669,24
mai/2016	14,15	0,78	11,2	42,20	672,08
jun/2016	14,15	0,35	11,3	41,87	702,89
jul/2016	14,15	0,52	11,6	42,11	700,36
ago/2016	14,15	0,44	11,8	42,17	696,75
set/2016	14,15	0,08	11,8	42,87	701,31
out/2016	14,05	0,26	11,8	43,08	669,38
nov/2016	13,90	0,18	11,9	43,15	671,13
dez/2016	13,65	0,30	12,0	41,97	671,13

Foram utilizados indicadores econômicos apurados no período de janeiro de 2015 a dezembro de 2016, obtidos no Sistema Gerenciador de Séries Temporais (SGS)[1].

A Tabela 4.1 apresenta os indicadores econômicos utilizados neste estudo e sua respectiva série histórica. O indicador Selic representa a taxa básica de juros definida pelo Comitê de Política Monetária do BACEN (COPOM) vigente no último dia do mês em observação e identificado pelo código 4189 no SGS. Os números 433, 24.369, 20.716 e 206 representam os códigos dos respectivos indicadores no SGS. O indicador Taxa Média de Juros é a média observada nas taxas praticadas pelos bancos brasileiros em operações de crédito para pessoas físicas ao longo de cada mês observado. Maiores detalhes sobre os demais indicadores podem ser consultados no SGS pesquisando pelo código identificado no cabeçalho da Tabela 4.1.

[1]https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries

Tabela 4.2: Tabela de Descrição das Variáveis Contínuas

Descrição
Idade do cliente
Número de dias em atraso do contrato.
Número de dias restantes para o fim do contrato.
Percentual de perda prevista para o contrato.
Quantidade de produtos possuídos pelo cliente.
Tempo de relacionamento do cliente com o Banco.
Valor da margem de contribuição total do cliente.
Valor da Renda Mensal do cliente.
Valor do PIB per capta do município de residência do cliente.
Valor do saldo devedor do contrato.
Valor original do contrato.
Valor provisionado para o contrato.

Variável resposta - Os dados extraídos do *data mart* tiveram sua atividade de preparação de dados facilitada, uma vez que a limpeza e padronização desses dados já havia sido realizada por ocasião de sua criação. Para a criação do rótulo da variável independente, o indicador de redução de atraso, foi realizada a seguinte operação, considerando que foram utilizados os dados das operações em atraso em um determinado mês:

- Com potencial de recuperação = 1, para todas as operações que apresentaram uma redução na quantidade de dias em atraso no mês subsequente ou que tiveram reduzidos os seus saldos devedores.

- Sem potencial de recuperação = 0, caso contrário, ou seja, apresentando manutenção ou aumento do atraso e saldo devedor maior ou igual ao mês anterior.

As Tabelas 4.2 e 4.3 apresentam as descrições das variáveis contínuas e categóricas que foram obtidas junto aos sistemas do Banco Alfa.

4.2 Entendimento dos Dados

A análise dos dados iniciou-se no mês de julho de 2016, contendo todos os contratos de operações de crédito, independente do produto contratado, com mais de 14 dias em atraso.

A Tabela 4.4 apresenta o resumo da extração das operações de PF em atraso no mês de julho de 2016, que resultou em uma base com 4.514.029 contratos. Deste total, apenas 271.193 (6.01%) apresentaram um potencial de recuperação.

Verificou-se que no Banco Alfa existem várias estratégias para a cobrança e recuperação de crédito, conforme o perfil do cliente e a categoria da operação de crédito, agrupando-os

Tabela 4.3: Tabela de descrição das variáveis categóricas

Descrição
Código do risco contábil da operação.
Estado do cadastro do cliente.
Estágio de relacionamento do cliente com o banco.
Faixa da estatística financeira da operação.
Faixa etária do cliente.
Faixa do tempo de relacionamento.
Gênero do cliente.
Indicador de operação estruturada.
Modalidade do produto contratado.
Natureza da ocupação do cliente.
Nível de instrução do cliente.
Nível gerencial que aprovou a operação.
Produto contratado.
Segmento comportamental do cliente.
Tipo da Carteira do cliente.
Tipo de trava de PCLD da operação.

Tabela 4.4: Composição da Base de Dados de Julho/2016

Sem potencial de recuperação	**Com Potencial de Recuperação**
4.242.836	271.193
93,99%	6,01%
Total de Registros	4.514.029

Tabela 4.5: Segmentos de Operações de Crédito

Segmento	**Potencial de Recuperação**				**Total de Registros**
	Não		**Sim**		
	Qtde	(%)	**Qtde**	(%)	
Negócios Sociais	137.474	93,53	9.504	6,47	146.978
Imobiliário I	41.398	70,45	17.365	29,55	58.763
Imobiliário II	400	73,94	141	26,06	541
Imobiliário III	3.537	78,11	991	21,89	4.528
Veículos I	12.115	87,90	1.667	12,10	13.782
Veículos II	32.357	86,63	4.993	13,37	37.350
Agronegócio	258.618	98,84	3.021	1,16	261.639
Cartão de Crédito I	17.124	98,92	187	1,08	17.311
Cartão de Crédito II	454.864	98,56	6.661	1,44	461,525
Demais Operações I	186.572	96,53	6.714	3,47	193.286
Demais Operações II	2.668.890	92,96	201.977	7,04	2.870.867

em segmentos com regras distintas de negociação. Os segmentos existentes são divididos em estratégias massificadas e individuais. As estratégias massificadas são implementadas para segmentos que possuem um padrão de comportamento conhecido, já as estratégias individuais abrangem operações que possuem características atípicas ou especiais, que necessitam de uma análise de cada caso para a realização de uma cobrança e recuperação.

Uma vez identificados esses segmentos de regras de negociação de operações em atraso, a base de dados foi separada em segmentos compatíveis com essas estratégias de recuperação da instituição, agrupando produtos similares e segmentos de clientes com características em comum, retirando do estudo os segmentos que possuem uma estratégia individualizada de negociação. A Tabela 4.5 lista os 11 segmentos que foram abordados neste estudo após a exclusão do segmento de Estratégia Individualizada, que representava 447.459 exemplares em julho de 2016.

Em seguida iniciou-se a preparação dos dados, analisando cada um dos segmentos, preparando-os para a fase de modelagem. O Banco Alfa autorizou a divulgação das análises descritivas desde que não fossem identificadas as variáveis, razão pela qual estas serão tratadas de forma anônima a partir deste ponto neste estudo. As variáveis identificadas no formato Vi são contínuas, enquanto VCi são categóricas.

Mesmo sem identificar as variáveis, a apresentação das análises descritivas justifica-se para evidenciar a inexistência de dados faltantes, a quantidade de classes das variáveis categóricas e o tau de Kendall[10] dessas variáveis em relação à variável resposta. O tau de Kendall foi calculado apenas para as variáveis numéricas e as categóricas ordinais.

Tabela 4.6: Imobiliário I - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	20	51	87,43	112	624	-0,29
V2	0	10.180	10.420	10.290	10.670	11.620	-0,03
V3	0	0,1	0,13	0,17	0,27	0,67	-0,02
V4	0	1	2	3,94	5	26	0,06
V5	17	25	29	31,28	36	73	0,03
V6	1	3	4	4,45	5	37	0,08
V7	14.790	74.400	87.460	86.270	97.470	164.800	0,01
V8	-124.400	-390,8	156,7	-1.397	256,3	183.400	0,38
V9	0	1.349	3.221	3.712	5.525	46.040	0,04
V10	0	888,1	2.670	19.090	20.960	173.700	-0,17
V11	0	1.586	1.700	1.877	2.000	20.000	0,03
V12	0,68	74.920	88.720	87.250	99.510	173.500	0,00

- Imobiliário I - O primeiro segmento analisado foi Imobiliário I. A Tabela 4.6 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável resposta. A Tabela 4.7 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Imobiliário II - A Tabela 4.8 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.9 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Imobiliário III - A Tabela 4.10 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.11 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Veículos I - A Tabela 4.12 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.13 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Veículos II - A Tabela 4.14 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.15 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

Tabela 4.7: Imobiliário I - Análise Descritiva das Variáveis Categóricas

Variável	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,23	6
VC2	0,02	5
VC3	-0,09	3
VC4	-0,21	9
VC5	0,03	12
VC6	0,06	7
VC7	-	2
VC8	-	5
VC9	-	16
VC10	0,01	4
VC11	0,05	8
VC13	-0,21	9
VC14	-	4
VC15	-	18
VC16	0,02	2

Tabela 4.8: Imobiliário II - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	21	48	89	113	507	-0,15
V2	0	1.626	2.928	3.037	4.076	6.776	-0,14
V3	0	0	0	0	0	1	0,09
V4	2	6	6	10	13	31	0,03
V5	30	42	48	49	55	85	0,02
V6	1	4	6	7	9	36	-0,04
V7	4.400	28.000	45.000	59.560	70.560	240.000	-0,05
V8	-31.770	-148	91	-650	371	25.070	0,30
V9	0	4.990	6.190	6.663	9.099	15.260	0,08
V10	0	173	650	8.744	10.230	116.000	-0,20
V11	0	1.598	2.965	5.082	5.553	128.900	-0,11
V12	0	9.316	20.500	36.670	47.120	215.200	-0,14

Tabela 4.9: Imobiliário II - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,11	4
VC2	0,13	3
VC4	-0,19	8
VC5	0,01	10
VC6	0,03	6
VC7	-0,04	2
VC8	0,10	5
VC9	-	10
VC11	-0,15	5
VC13	-0,19	8
VC14	0,10	4
VC15	-	13

Tabela 4.10: Imobiliário III - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	30	81	131	181	511	-0,31
V2	0	4.876	7.530	6.616	8.203	10.910	-0,01
V3	0,00	0,02	0,05	0,06	0,07	0,10	0,00
V4	0	5	9	11	15	54	-0,02
V5	20	35	43	44	52	78	-0,06
V6	2	6	8	10	11	71	0,05
V7	20.000	100.000	142.500	188.700	213.800	3.000.000	-0,07
V8	-257.600,00	-5.476,00	-930,00	-9.087,00	257,70	199.600	0,36
V9	0	2.769	4.660	4.968	6.485	46.040	0,01
V10	0	3.317	14.200	51.540	53.470	1.212.000	-0,20
V11	0	2.280	5.542	10.700	11.130	337.600	-0,01
V12	250	88.900	134.500	177.500	203.200	3.084.000	-0,08

Tabela 4.11: Imobiliário III - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,16	6
VC2	0,01	4
VC3	-0,24	2
VC4	-0,21	33
VC5	-0,06	12
VC6	-0,02	7
VC7	-	2
VC8	-	5
VC9	-	16
VC10	-0,10	5
VC11	0,03	7
VC12	-	2
VC13	-0,21	9
VC14	-	5
VC15	-	17
VC16	0,17	2

Tabela 4.12: Veículos I - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	46	98	131	193	711	-0,32
V2	0	375	710	733,3	1.075	1.710	-0,02
V3	0	0,19	0,255	0,27	0,352	1,44	-0,01
V4	0	4	10	11,12	16	60	0,08
V5	19	34	43	43,94	52	96	0,03
V6	1	8	11	12,69	15	183	0,04
V7	3.353	24.410	35.070	44.360	52.850	514.900	-0,03
V8	-1.593.000	-6.039	-628,9	-10.210	580,5	719.600	0,25
V9	0	2.133	3.717	4.384	5.838	46.040	0,00
V10	0	2.236	10.810	18.210	24.180	327.700	-0,23
V11	291,7	9.249	12.600	24.270	21.750	1.116.000	-0,03
V12	1,44	14.210	24.590	33.180	41.540	5.27.200	-0,09

Tabela 4.13: Veículos I - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,33	4
VC2	-0,01	4
VC3	0,08	2
VC4	-0,24	18
VC5	0,03	16
VC6	0,09	7
VC7	-	2
VC8	-	8
VC9	-	15
VC11	0,03	8
VC12	-	2
VC13	-0,25	9
VC14	-	2
VC15	-	15
VC16	-0,08	2

Tabela 4.14: Veículos II - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	43	104,5	140,2	208	718	-0,36
V2	0	314	559	627,8	894	1.746	0,01
V3	0	0,27	0,42	0,41	0,50	1,76	-0,04
V4	0	3	7	8,36	12	54	0,07
V5	18	31	38	39,57	47	96	0,03
V6	1	5	7	7,46	9	50	0,07
V7	2.033	13.210	19.370	22.420	27.770	312.800	0,01
V8	-151.700	-632,8	-69,47	-910,3	331,3	215.200	0,27
V9	0	1.550	3.204	3.815	5.174	46.040	-0,01
V10	0,01	457,6	4.275	7.796	11.640	168.500	-0,26
V11	0	1.820	3.200	3.529	4.940	25.440	0,03
V12	0,01	7.508	12.330	15.210	19.430	196.200	-0,07

- Agronegócios - A Tabela 4.16 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.17 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

Tabela 4.15: Veículos II - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,24	7
VC2	-0,01	5
VC3	0,15	2
VC4	-0,29	16
VC5	0,03	16
VC6	0,07	7
VC7	-	2
VC8	-	7
VC9	-	19
VC11	0,02	8
VC12	-	2
VC13	-0,30	8
VC14	-	3
VC15	-	18
VC16	-0,15	2

Tabela 4.16: Agronegócios - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	97	183	204,3	266	4.256	-0,06
V2	0	0	390	885,7	1.662	4.702	0,06
V3	0	0,03	0,05	0,18	0,11	8,644	0,01
V4	0	5	9	10,43	14	68	0,02
V5	18	35	46	46,95	57	104	0,01
V6	1	3	6	8,80	11	207	-0,02
V7	90	9.000	18.710	54.970	47.140	14.550.000	-0,04
V8	-2.410.000	-1.004	28,75	-5.911	159	719.600	0,10
V9	0	883,6	2.132	2.527	3.688	46.040	-0,01
V10	-9,73	620,5	3.065	19.230	11.870	4.853.000	-0,04
V11	0	1.025	2.288	11.230	6.533	1.211.000	-0,03
V12	0,01	2.028	5.899	24.750	17.080	5.150.000	0,01

Tabela 4.17: Agronegócios - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,06	7
VC2	0,01	6
VC3	0,01	4
VC4	-0,07	26
VC5	0,01	16
VC6	0,02	7
VC7	-	2
VC9	-	19
VC10	-	7
VC11	-0,02	8
VC12	-	22
VC13	-0,07	9
VC14	-	5
VC15	-	21
VC16	0,11	2

- Negócios Sociais - A Tabela 4.18 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.19 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Cartões de Crédito I - A Tabela 4.20 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.21 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Cartões de Crédito II - A Tabela 4.22 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.23 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Demais Operações I - A Tabela 4.24 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à variável de interesse. A Tabela 4.25 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

- Demais Operações II - A Tabela 4.26 apresenta o resumo da análise descritiva das variáveis contínuas e o cálculo do tau de Kendall de cada variável com relação à

Tabela 4.18: Negócios Sociais - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	71	141	171,9	259	720	-0,31
V2	0	0	0	123,3	177	1.780	0,26
V3	0	0,52	0,64	0,96	1,28	3,24	0,04
V4	0	3	4	6,22	9	45	0,04
V5	17	28	37	38,19	47	93	0,03
V6	1	5	6	6,81	8	50	0,01
V7	101	1.339	2.060	2.502	3.090	30.810	-0,04
V8	-350.200	-206	-20,9	-249,2	82,32	118.100	0,20
V9	0	1.610	3381	4.080	5.667	46.040	-0,02
V10	0	48,05	316,7	804,8	1.091	30.140	-0,20
V11	0	1.500	1.700	1.885	1.860	79.680	0,01
V12	0,03	766,7	1.303	1.715	2.127	31.170	-0,03

Tabela 4.19: Negócios Sociais - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,22	7
VC2	0,01	5
VC3	0,16	2
VC4	-0,24	14
VC5	0,03	16
VC6	0,04	7
VC7	-	2
VC8	-	5
VC9	-	20
VC11	-0,01	8
VC12	-	4
VC13	-0,26	7
VC14	-	4
VC15	-	19
VC16	-0,16	2

Tabela 4.20: Cartão de Crédito I - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	25	63	84,54	102	706	-0,04
V2	0	0	5	5,445	10	13	-0,12
V3	0,098	0,219	0,275	0,3114	0,333	3,803	0,03
V4	0	7	13	13,09	18	63	0,01
V5	18	36	45	46,6	55	101	0,01
V6	1	8	11	12,92	17	108	-0,03
V7	1	1900	5000	7461	9997	100000	-0,02
V8	-595700	-1830	85,37	-4051	1308	302200	0,04
V9	0	2684	4658	4984	7897	46040	-0,01
V10	0	9,51	47,7	731,6	325,9	38030	-0,05
V11	0	8749	10490	13930	14170	1116000	-0,02
V12	0,1	189,2	1246	3255	3760	83740	-0,05

Tabela 4.21: Cartão de Crédito I - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,15	6
VC2	0,01	6
VC3	0,04	2
VC4	-0,04	18
VC5	0,01	16
VC6	-0,01	7
VC7	-	3
VC8	-	51
VC9	-	19
VC10	0,03	2
VC11	-	8
VC13	-0,03	9
VC14	-	2
VC15	-	18
VC16	-0,04	2

Tabela 4.22: Cartão de Crédito II - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	25	71	84,42	112	720	-0,04
V2	0	0	6	5,86	10	13	-0,14
V3	0	0,43	0,55	0,66	0,67	6,72	0,02
V4	0	3	6	7,26	11	63	0,02
V5	8	26	34	37,4	46	116	0,02
V6	1	4	5	6,17	8	83	0,02
V7	1	326	550	1.248	1.167	10.000.000	0,01
V8	-116.600	-147,80	2,20	-201,9	110,90	85.890	0,09
V9	0	1.962	3.648	4.373	5.976	46.040	-0,02
V10	-23,39	2,33	11,80	143,3	87,46	44.120	-0,05
V11	0	1.020	1.646	1.968	2.441	47.170	0,03
V12	0,01	50,26	226,40	619,30	682	64.350	-0,04

Tabela 4.23: Cartão de Crédito II - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,04	8
VC2	0,01	6
VC3	0,01	2
VC4	-0,04	18
VC5	0,02	18
VC6	0,02	7
VC7	-	3
VC8	-	65
VC9	-	22
VC10	0,03	2
VC11	-	8
VC13	-0,04	9
VC13	-	3
VC15	-	20
VC16	-0,01	2

Tabela 4.24: Demais Operações I - Análise Descritiva das Variáveis Contínuas

Variável	**Min**	**1QT**	**Mediana**	**Média**	**3QT**	**Máximo**	**Kendall**
V1	15	89	167	186,3	269	720	-0,23
V2	0	84	472	696,6	1.212	2.866	0,03
V3	0	0,36	0,52	0,53	0,64	16,33	0,02
V4	0	6	12	12,91	18	76	0,01
V5	15	37	46	47,58	56	107	0,01
V6	1	9	13	15,45	18	156	-0,01
V7	29,98	2.246	6.300	15.800	17.000	754.400	-0,02
V8	-2.410.000	-2.944	-407,4	-6.828	402,5	886.500	0,15
V9	0	2.684	4.565	4.919	74.86	46.040	-0,04
V10	-5.530	354,6	2.012	8.409	8.022	357.100	-0,15
V11	0	9.000	11.460	18.050	16.930	1.116.000	-0,02
V12	0,03	2.376	7.187	18.330	20.090	850.700	-0,05

Tabela 4.25: Demais Operações I - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,34	6
VC2	0,01	6
VC3	0,10	3
VC4	-0,18	18
VC5	0,01	16
VC6	0,02	7
VC7	-	3
VC8	-	28
VC9	-	19
VC10	0,02	8
VC11	-	8
VC12	0,01	6
VC13	-0,19	9
VC14	-	2
VC15	-	17
VC16	-0,09	2

Tabela 4.26: Demais Operações II - Análise Descritiva das Variáveis Contínuas

Variável	Min	1QT	Mediana	Média	3QT	Máximo	Kendall
V1	15	69	147	167,8	248	720	-0,31
V2	0	34	239	425,9	583	2.876	0,08
V3	0	0,67	1,00	1,22	1,59	16,33	0,03
V4	0	3	6	7,48	11	64	0,05
V5	4	28	36	39,16	48	113	0,04
V6	1	5	6	7,54	9	117	0,04
V7	2,23	400	930,1	2.799	2.400	807.100	-0,01
V8	-326.100	-218,8	-30,23	-424,9	43,34	160.000	0,25
V9	0	1.763	3.462	4.129	5.667	46.040	-0,04
V10	-7.444	44,14	275,60	1.279	997,8	239.900	-0,23
V11	0	1.199	1.750	2.220	2.709	50.470	0,01
V12	0,01	437,5	1.037	2.885	2.537	436.200	-0,06

variável de interesse. A Tabela 4.27 apresenta o resumo da análise descritiva das variáveis categóricas, o cálculo do tau de Kendall e a quantidade de classes.

4.3 Preparação de Dados

A criação dos 11 segmentos listados na Tabela 4.5 seguiu a estratégia de recuperação atual do Banco Alfa, baseada na combinação de produtos de crédito e o nível de renda dos clientes. Nesse estudo ampliamos essa separação, dividindo cada um dos segmentos em três subsegmentos, conforme a origem da renda dos clientes, a saber: setor público, setor privado e renda incerta. Neste último, foram incluídos os clientes que não possuíam uma renda formal ou regular, como autônomos, bolsistas e rendas temporárias.

A Tabela 4.28 representa a criação dos subsegmentos para o segmento Demais Operações I. Esta regra foi aplicada a todos os segmentos.

4.3.1 Bases de Treinamento e Validação

As bases de treinamento e validação foram divididas de modo a permitir a realização de duas estratégias de modelagem para cada um dos segmentos, uma anual e outra mensal.

Modelagem Anual

Para a modelagem anual, o treinamento foi formado pelo conjunto de operações dos meses de janeiro a dezembro de 2015, formando uma única base para cada um dos segmentos e subsegmentos.

Tabela 4.27: Demais Operações II - Análise Descritiva das Variáveis Categóricas

Variavel	**Tau de Kendall**	**Quantidade de Classes**
VC1	0,12	8
VC2	0,02	6
VC3	0,10	3
VC4	-0,27	18
VC5	0,04	18
VC6	0,06	7
VC7	-	3
VC8	-	31
VC9	-	22
VC10	0,01	8
VC11	-	9
VC12	0,01	6
VC13	-0,28	9
VC14	-	3
VC15	-	21
VC16	-0,09	2

Tabela 4.28: Exemplo da criação dos subsegmentos

Segmento	Subsegmento
Demais Operações I	Setor Público
	Setor Privado
	Renda Incerta

Modelagem Mensal

Já na modelagem mensal, foram criadas 12 bases de treinamento, cada uma referente a um mês do ano de 2015, para cada um dos segmentos e subsegmentos.

Validação

Os modelos gerados com a visão anual e mensal foram validados sob a mesma base de validação para permitir a comparação de performance entre eles. Foram criadas 12 bases de validação para cada um dos segmentos e subsegmentos, relativos aos meses de janeiro a dezembro de 2016.

Capítulo 5

Modelagem e Validação

Neste capítulo são detalhados os procedimentos de modelagem e suas respectivas validações, dividido em seções conforme os segmentos estudados nesta pesquisa.

Para cada um dos segmentos e subsegmentos, foram elaborados modelos preditivos com os algoritmos GLM, GBM, DRF e DL[1] na plataforma H2O.ai por meio de chamadas a partir de um código R. Os três primeiros algoritmos foram escolhidos por representarem as técnicas mais utilizadas no cálculo de risco de crédito, que realiza uma tarefa de classificação muito similar ao presente estudo[5]. Já o algoritmo DL foi utilizado para verificação do seu comportamento em uma área de conhecimento ainda não explorada, mas com expectativa de boa adequação, em virtude da utilização de uma grande quantidade de variáveis [4].

5.1 Ajustes de Parâmetros

Os parâmetros de cada modelo foram ajustados por meio da utilização da funcionalidade *grid search* da plataforma H2O.ai, onde é possível estabelecer uma série de hiperparâmetros e os respectivos valores a serem testados. A partir da combinação cartesiana desses parâmetros, foram realizados os treinamentos dos modelos em busca da maior média harmônica entre a precisão e a sensibilidade, chamada de f1-*score*.

Os hiperparâmetros e os seus respectivos valores estão listados na Tabela 5.1. O parâmetro *ntree* define a quantidade máxima de árvores de decisão elaboradas no modelo, enquanto *max_depth* é o limite de profundidade de cada árvore de decisão. O parâmetro *balance_classes* promove o balanceamento de classes na base de treinamento, promovendo o aumento de exemplares da classe minoritária, conhecido como *oversampling*. O problema de balanceamento de classes afeta não apenas as árvores de decisão como também as

[1] *Deep Learning* é o nome dado à implementação do algoritmo de Redes Neurais Artificias com múltiplas camadas na plataforma H2O.ai

Tabela 5.1: Tabela de Hiperparâmetros

Algoritmo	Hiperparâmetros
DRF e GBM	*ntree*: 200, 400 e 600
	max_depth: 7, 10, 13
	balance_classes: true
GLM	*alpha*: 0,4 , 0,5 e 0,6
	lambda: 1e-05, 1e-04 e 1e-03
DL	*hidden*: (50,50). (100,100), (150,150),
	(50,50,50). (100,100,50), (150,150,50)
	epochs: 5, 10, 15, 20 e 25
	balance_classes: true

redes neurais, ainda que em menor impacto, e a técnica de *oversampling* pode melhorar a performance do modelo [11]. [12][2]

O *alpha* e o *lambda* funcionam do modo combinado, promovendo a regularização do modelo, na tentativa de evitar-se um superajuste do modelo (*overfitting*). Duas formas muito utilizadas na regularização de modelos lineares generalizados são o *Least Absolute Shrinkage and Selection Operator* (LASSO) e *Ridge Regression*[13], o parâmetro *alpha* permite a combinação da utilização dessas duas técnicas, indicando o peso de utilização entre elas. Todos os modelos treinados com GLM tiveram as variáveis numéricas normalizadas pelo próprio algoritmo que, por padrão, realiza essa transformação.[3]

Os parâmetros *hidden* e *epochs* representam a configuração da camada oculta e a quantidade máxima de épocas de treinamento das redes neurais.

Para todos os segmentos, a escolha do algoritmo e de seus parâmetros foi realizada pela observação do melhor *f1-score* obtido com o treinamento do mês de julho de 2016 e a validação cruzada com dez *folds*. O algoritmo vencedor de cada segmento, e seus respectivos parâmetros, foram utilizados na modelagem segundo a perspectiva mensal e anual.

5.2 Negócios Sociais

Esta seção apresenta com detalhes as atividades realizadas na modelagem do segmento Negócios Sociais e seus subsegmentos Setor Público, Setor Privado e Renda Incerta.

A Tabela 5.2 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento.

[2]O parâmetro *balance_classes* consta na Tabela 5.1 apenas para evidenciar o uso do balanceamento (*oversampling*)

[3]Disponível em http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/standardize.html. Acessado em junho de 2016.

Tabela 5.2: Negócios Sociais - Parâmetros do Algoritmo

Segmento	Subsegmento	Algoritmo	Parâmetros
Negócios Sociais	Total	GBM	ntrees = 600, max_depth = 13

Tabela 5.3: Validações do Segmento Negócios Sociais - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	69	78	69	70	67	73	79	80	74	76	75	83
Precisão	81	70	95	82	97	95	87	83	88	86	80	79
Acurácia	71	77	77	72	76	79	80	80	76	78	76	83

5.2.1 Modelagem Anual

Inicialmente, a modelagem foi realizada considerando apenas o segmento, ou seja, utilizando todas as operações dos clientes independentes do segmento ao qual pertenciam (setor público, privado ou renda incerta).

A Tabela 5.3 apresenta os indicadores observados nas 12 validações realizadas ao longo do ano de 2016. A especificidade, que mede a taxa de acerto da classe negativa, ou seja, os clientes sem potencial de recuperação, variou entre 67% e 80%. Já a precisão, que mede a taxa de acerto da classe positiva, apresentou uma variação entre 79% e 97%. A acurácia, que mede a taxa de acerto total, ou seja, das classes negativas e positivas, variou de 71% a 83%.

A Figura 5.1 apresenta os *boxplots*[14] dos indicadores observados no segmento Negócios Sociais. As métricas P0, P1, PT representam os indicadores Especificidade, Precisão e Acurácia, respectivamente.

Avaliação dos Subsegmentos

Em seguida, foram treinados os modelos para os subsegmentos, utilizando-se os algoritmos e parâmetros listados na Tabela 5.4

A Tabela 5.5 apresenta os indicadores obtidos nas validações realizadas nos modelos treinados com cada subsegmento de Negócios Sociais. Estes subsegmentos apresentaram distintos entre si e também em relação ao modelo treinado com todos os subsegmentos.

Tabela 5.4: Negócios Sociais - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Negócios Sociais	Setor Público	GBM	ntrees = 200, max_depth = 10
	Setor Privado	GBM	ntrees = 600, max_depth = 13
	Renda Incerta	GBM	ntrees = 400, max_depth = 13

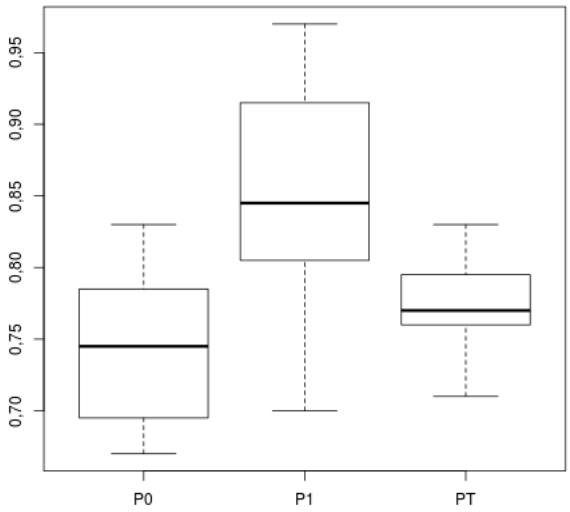

Figura 5.1: Negócios Sociais - Modelagem Anual

Tabela 5.5: Resultado das Validações dos Subsegmentos de Negócios Sociais - Anual

2016													
Público	Especificidade	45	68	71	51	60	60	69	67	67	69	63	73
	Precisão	95	74	90	90	94	98	89	90	86	80	83	89
	Acurácia	58	69	80	61	75	76	73	71	71	71	69	75
Privado	Especificidade	69	75	86	70	69	77	80	80	76	79	81	80
	Precisão	87	78	76	87	96	91	90	86	89	83	72	88
	Acurácia	72	75	83	73	76	81	81	80	77	80	79	81
Renda Incerta	Especificidade	67	73	74	72	68	75	80	85	78	76	69	80
	Precisão	88	81	91	82	96	93	88	73	89	91	88	91
	Acurácia	70	75	79	74	76	80	81	83	79	78	72	81

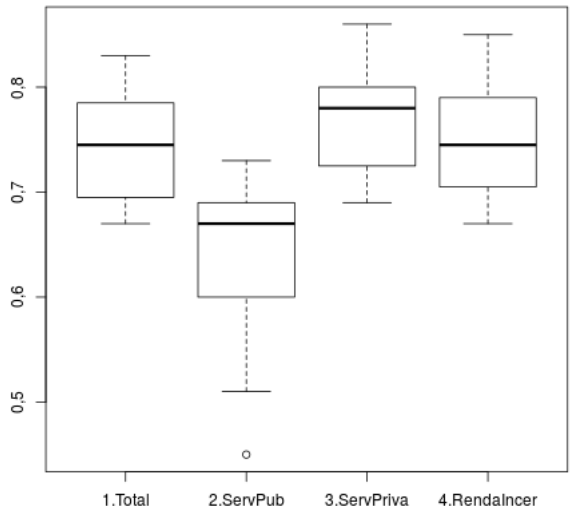

Figura 5.2: Negócios Sociais - Modelagem Anual - Especificidade

Para comparar a performance dos modelos foram gerados *boxplots* para cada indicador, de modo que pudesse ser observado o resultado de cada subsegmento.

A Figura 5.2 apresenta a comparação da Especificidade, obtido pelas 12 validações realizadas em cada um dos subsegmentos e do segmento total. É possível perceber que os resultados observados no subsegmento Setor Público apresentam um comportamento bem diferenciado em relação aos demais. Foi o subsegmento que apresentou a menor performance de especificidade e, por representar apenas cerca de 18% do segmento, não teve forte influência no resultado da avaliação global do segmento.

Já a Figura 5.3 apresenta a comparação do indicador Precisão e, desta vez, demonstra resultados semelhantes, uma vez que é possível observar uma faixa de resultados comum a todos subsegmentos.

Enfim, a Figura 5.4 ilustra o comparativo do indicador Acurácia, permitindo uma avaliação global da performance dos modelos em cada um dos subsegmentos e o segmento em sua perspectiva global.

Escolha entre Segmento ou Subsegmentos

O comparativo das performances obtidas pelos três indicadores mostrou que, em todos os casos, o subsegmento Setor Privado apresentou um comportamento bem distinto dos demais. Por essa razão, a estratégia de segregação do segmento em subsegmentos demonstrou ser correta para Negócios Sociais, na visão anual.

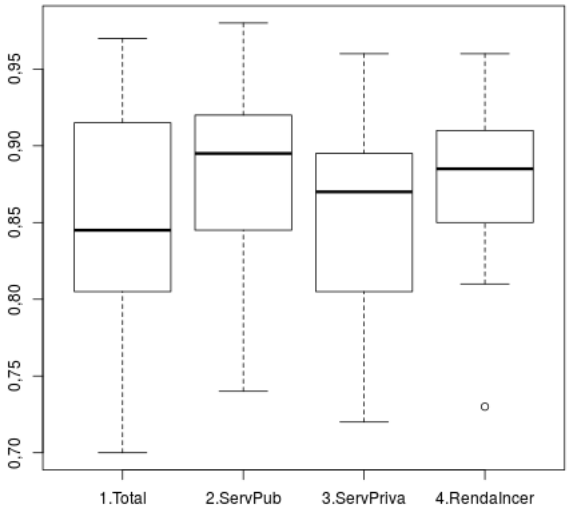

Figura 5.3: Negócios Sociais - Modelagem Anual - Precisão

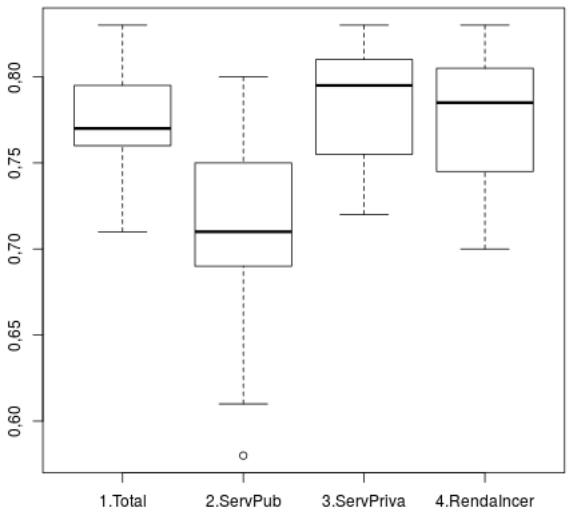

Figura 5.4: Negócios Sociais - Modelagem Anual - Acurácia

Tabela 5.6: Validações do Segmento Negócios Sociais - Mensal

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	84	87	93	87	85	92	97	97	90	93	96	98
Precisão	79	69	77	76	90	86	64	60	79	75	67	53
Acurácia	83	85	89	85	86	91	95	95	89	91	93	95

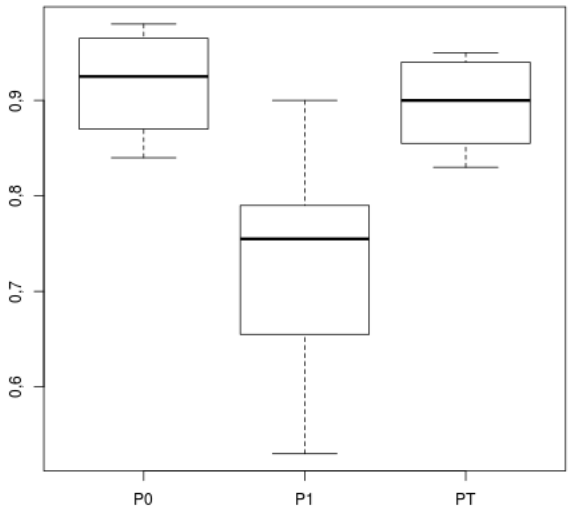

Figura 5.5: Negócios Sociais - Modelagem Mensal

5.2.2 Modelagem Mensal

Para a modelagem mensal, seguiu-se a mesma estratégia de avaliação do segmento Negócios Sociais comparada aos seus subsegmentos. Porém, neste caso, foram treinados modelos para cada mês de observação em 2015 e validou-se com o respectivo mês do ano de 2016. Ou seja, para o modelo treinado com os dados de janeiro de 2015, a validação foi realizada frente aos dados de janeiro de 2016. O treinamento de fevereiro de 2015 foi validado com fevereiro de 2016 e, sucessivamente, até a validação do mês de dezembro de 2016.

A Tabela 5.6 apresenta as validações realizadas em cada um dos modelos treinados e seus respectivos indicadores. A especificidade variou entre 84% e 97%, a precisão apresentou uma variação entre 60% e 90% e, por fim, a acurácia variou de 83% a 95%.

A Figura 5.5 apresenta os *boxplots* dos indicadores observados no segmento Negócios Sociais na modelagem mensal.

Tabela 5.7: Resultado das Validações dos Subsegmentos de Negócios Sociais - Mensal

		2016											
	Indicador	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Público	Especificidade	68	67	81	69	74	81	88	84	80	85	91	87
	Precisão	89	88	89	88	94	93	78	87	85	78	74	82
	Acurácia	73	72	84	72	79	84	87	84	81	84	88	86
Privado	Especificidade	86	88	92	88	85	92	96	95	88	94	94	98
	Precisão	76	67	80	71	79	80	65	69	86	75	76	52
	Acurácia	85	86	90	86	84	91	94	94	88	92	93	96
Renda Incerta	Especificidade	88	88	94	92	84	93	98	98	90	95	97	98
	Precisão	69	72	73	62	91	82	58	53	84	64	61	55
	Acurácia	86	85	90	89	85	92	96	96	89	93	94	96

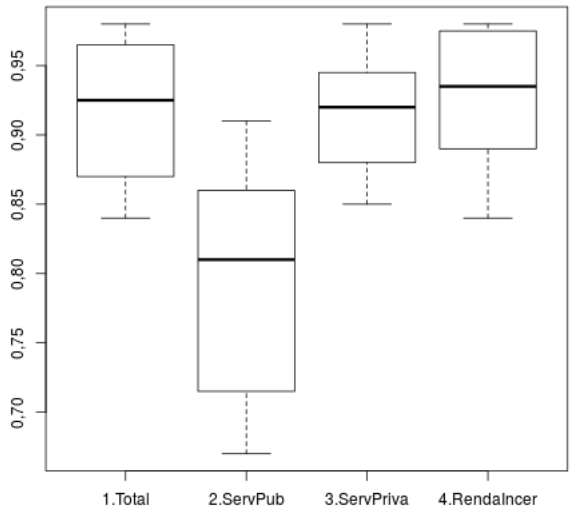

Figura 5.6: Negócios Sociais - Modelagem Mensal - Especificidade

Avaliação dos Subsegmentos

Assim como na modelagem anual, foram treinados modelos para cada um dos subsegmentos, utilizando-se os algoritmos e parâmetros listados na Tabela 5.4

A Tabela 5.7 apresenta os resultados obtidos nas validações dos modelos mensais em cada subsegmento de Negócios Sociais. Assim como na modelagem anual, estes subsegmentos apresentaram distintos entre si e o segmento completo.

A Figura 5.6 apresenta a comparação da especificidade em cada um dos subsegmentos e do segmento total. Os resultados foram semelhantes aos observados na modelagem anual, no sentido de que um dos subsegmentos, o Setor Público, também apresentou um comportamento distinto dos demais.

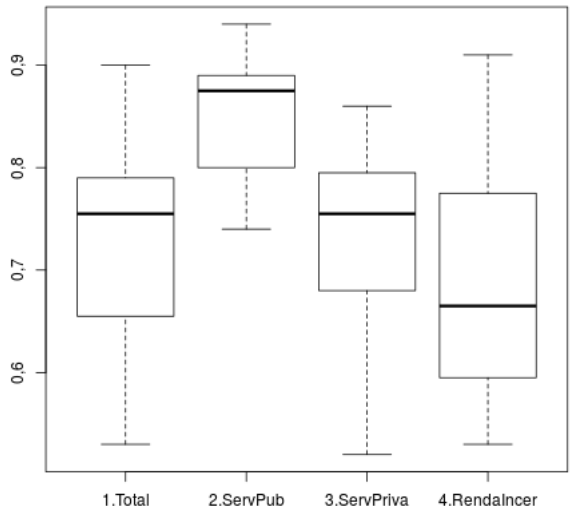

Figura 5.7: Negócios Sociais - Modelagem Mensal - Precisão

A Figura 5.7 apresenta a comparação do indicador Precisão entre os subsegmentos e confirma-se novamente o comportamento diferenciado do subsegmento Setor Privado.

Finalmente, a Figura 5.8 apresenta o comparativo do indicador Acurácia diante dos subsegmentos e o segmento total de Negócios Sociais.

Escolha entre Segmento ou Subsegmentos

Foi utilizado o mesmo critério de decisão da análise da modelagem anual, optando-se pela escolha dos modelos treinados por subsegmento, em razão da comprovação do comportamento distinto entre o segmento completo e seus subsegmentos.

5.2.3 Comparativo da Modelagem Anual e Mensal

Como as classes negativas, clientes sem potencial de recuperação, e as classes positivas, clientes com potencial de recuperação, são de interesse para realização da estratégia de recuperação de operações de crédito do Banco, optou-se por tomar o indicador acurácia como referência para a escolha da melhor modelagem a ser utilizada, anual ou mensal.

A Figura 5.9 compara os resultados da acurácia da modelagem anual e mensal, permitindo-nos observar que, na maior parte das validações, a modelagem mensal apresentou uma performance superior à modelagem anual.

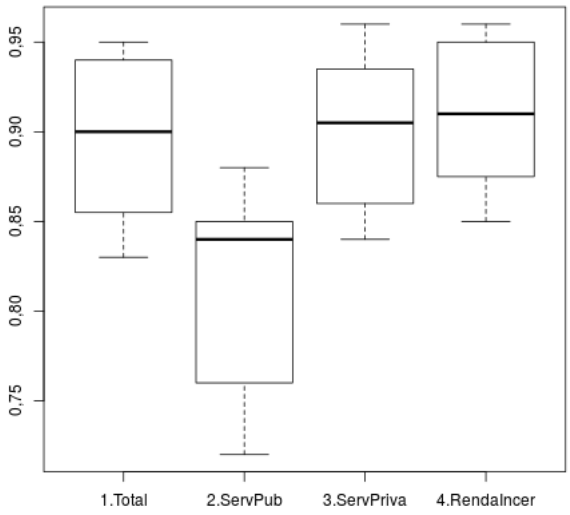

Figura 5.8: Negócios Sociais - Modelagem Anual - Acurácia

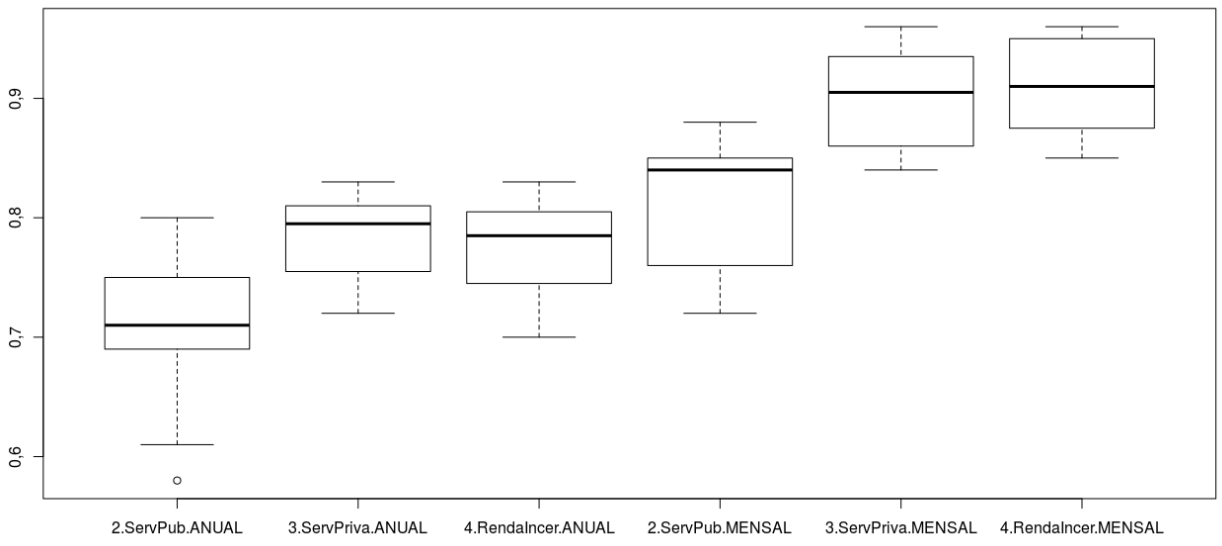

Figura 5.9: Negócios Sociais - Modelagem Anual - Acurácia

Tabela 5.8: Imobiliário I - Parâmetros do Algoritmo

Segmento	Subsegmento	Algoritmo	Parâmetros
Imobiliário I	Total	DL	hidden = (150, 150, 50), epochs = 20

Tabela 5.9: Validações do Segmento Imobiliário I - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	52	40	71	53	46	42	69	69	66	51	28	71
Precisão	84	97	87	88	93	95	79	72	78	86	96	59
Acurácia	59	54	78	63	69	64	72	70	69	60	52	69

5.3 Imobiliário I

Esta seção apresentará os procedimentos de modelagem, validação e seleção de modelos para o segmento Imobiliário I e os seus subsegmentos.

A Tabela 5.8 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento.

5.3.1 Modelagem Anual

Assim como realizado no estudo do segmento Negócios Sociais, a modelagem iniciou abordando o segmento completo, ou seja, utilizando todas as operações dos clientes independente do segmento ao qual pertenciam (setor público, privado ou renda incerta).

A Tabela 5.9 apresenta a apuração da validação realizada no segmento Imobiliário I. A especificidade variou entre 28% e 71%. Já a precisão apresentou uma variação entre 59% e 97%. A acurácia variou de 52% a 78%.

A Figura 5.10 apresenta os *boxplots* dos indicadores observados no segmento Imobiliário I. As métricas P0, P1, PT representam os indicadores Especificidade, Precisão e Acurácia, respectivamente.

Avaliação dos Subsegmentos

Em seguida, foram treinados os modelos para os subsegmentos, utilizando-se os algoritmos e parâmetros listados na Tabela 5.10

Tabela 5.10: Imobiliário I - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Imobiliário	Setor Público	GBM	ntrees = 200, max_depth = 13
	Setor Privado	GBM	ntrees = 200, max_depth = 13
	Renda Incerta	GBM	ntrees = 600, max_depth = 13

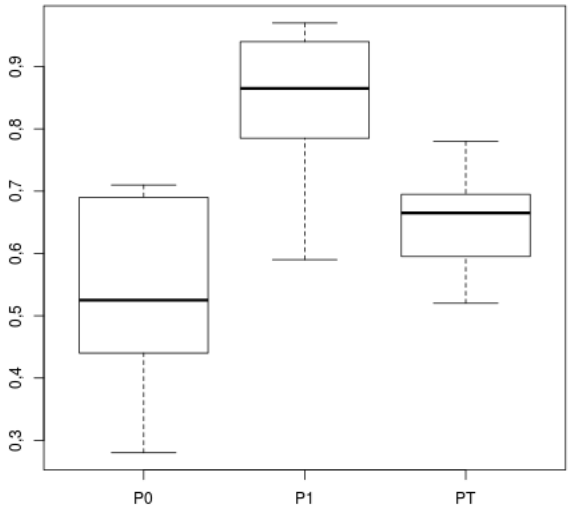

Figura 5.10: Imobiliário I - Modelagem Anual

Tabela 5.11: Resultado das Validações dos Subsegmentos de Imobiliário I - Anual

		2016											
	Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Público	Especificidade	47	52	43	43	44	44	47	58	66	54	33	63
	Precisão	89	84	98	93	95	98	96	85	78	84	97	82
	Acurácia	58	63	71	61	71	70	64	66	69	64	64	67
Privado	Especificidade	48	53	58	49	43	51	61	56	52	47	32	57
	Precisão	86	85	93	90	96	95	89	84	89	90	96	82
	Acurácia	56	62	74	61	70	71	70	63	62	59	57	61
Renda Incerta	Especificidade	52	67	66	58	52	56	62	56	59	67	43	66
	Precisão	88	75	91	86	92	94	94	93	90	75	89	74
	Acurácia	59	68	76	66	71	71	70	64	66	69	58	67

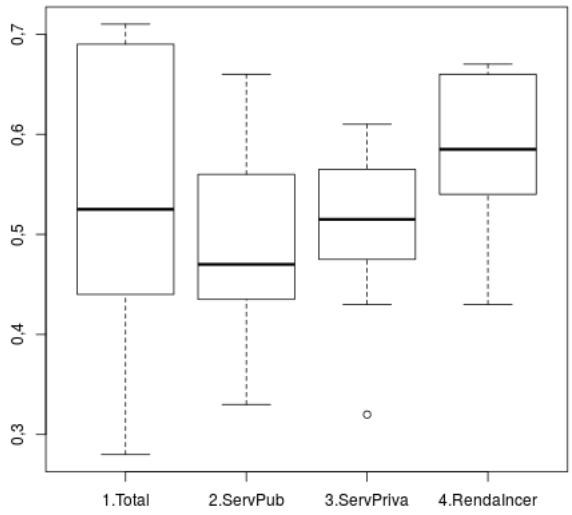

Figura 5.11: Imobiliário I - Modelagem Anual - Especificidade

A Tabela 5.11 apresenta os indicadores obtidos nas validações realizadas nos modelos treinados com cada subsegmento de Imobiliário I. Estes subsegmentos apresentaram distintos entre si e também em relação ao modelo treinado com todos os subsegmentos.

A Figura 5.11 apresenta a comparação do Especificidade, obtida pelas 12 validações realizadas em cada um dos subsegmentos e do segmento total. É possível perceber que os resultados observados possuem regiões coincidentes entre os subsegmentos e segmento completo.

Já a Figura 5.3 indica a comparação do indicador Precisão e, novamente, apresentam resultados semelhantes, uma vez que é possível observar uma faixa de resultados comum a todos os subsegmentos e o segmento completo.

Enfim, a Figura 5.13 ilustra o comparativo do indicador Acurácia, permitindo uma avaliação global da performance dos modelos em cada um dos subsegmentos e o segmento em sua perspectiva global.

Escolha entre Segmento ou Subsegmentos

O comparativo das performances obtidas pelos três indicadores mostrou que, em todos os casos, os *boxplots* apresentaram uma faixa de resultados coincidentes, que mostra que a utilização de qualquer uma das duas estratégias geraria resultados semelhantes. Por essa razão, a estratégia de utilização do segmento completo foi escolhida, diminuindo a quantidade de modelos gerados, privilegiando um modelo mais generalizado.

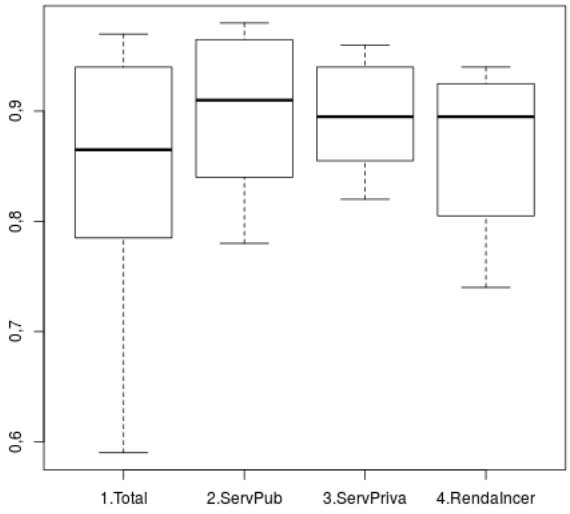

Figura 5.12: Imobiliário I - Modelagem Anual - Precisão

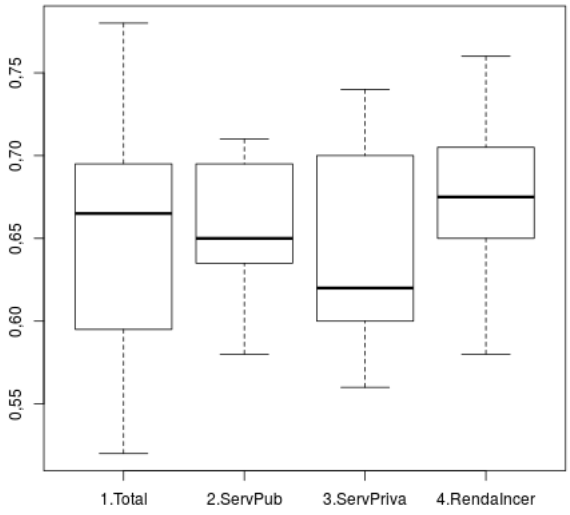

Figura 5.13: Imobiliário I - Modelagem Anual - Acurácia

Tabela 5.12: Validações do Segmento Imobiliário I - Mensal

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	61	56	58	53	56	49	64	59	34	53	36	51
Precisão	75	84	88	82	80	94	87	85	92	88	92	87
Acurácia	64	63	71	62	68	68	71	66	48	63	56	57

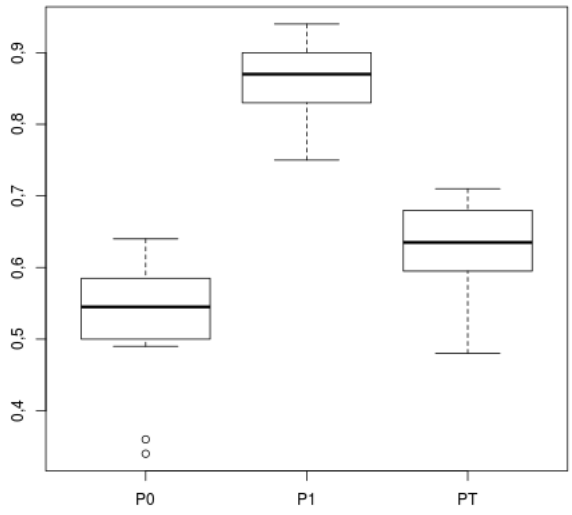

Figura 5.14: Imobiliário I - Modelagem Mensal

5.3.2 Modelagem Mensal

Para a modelagem mensal do segmento Imobiliário I, seguiu-se a mesma estratégia utilizada no segmento Negócios Sociais treinando modelos conforme cada mês de 2015 e validando-se com o respectivo mês do ano de 2016.

A Tabela 5.12 exibe as validações realizadas em cada um dos modelos treinados e seus respectivos indicadores. A especificidade variou entre 34% e 61%, a precisão apresentou uma variação entre 72% e 94% e, por fim, a acurácia variou de 48% a 71%.

A Figura 5.14 apresenta os *boxplots* dos indicadores observados no segmento Imobiliário I na modelagem mensal.

Avaliação dos Subsegmentos

Assim como na modelagem anual, foram treinados modelos para cada um dos subsegmentos, utilizando-se os algoritmos e parâmetros listados na Tabela 5.10

Tabela 5.13: Resultado das Validações dos Subsegmentos de Imobiliário I - Mensal

		2016											
	Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Público	Especificidade	42	49	43	39	46	41	49	52	42	47	29	59
	Precisão	90	84	96	95	89	98	94	88	95	90	97	79
	Acurácia	55	61	70	59	68	68	64	62	59	61	62	63
Privado	Especificidade	47	39	51	40	40	47	60	53	50	48	33	56
	Precisão	85	93	94	92	93	97	89	87	88	90	96	83
	Acurácia	56	54	70	55	67	69	69	62	61	60	57	60
Renda Incerta	Especificidade	56	58	58	49	41	52	64	57	52	55	39	73
	Precisão	80	80	90	88	93	96	90	93	94	89	95	65
	Acurácia	60	63	71	59	66	69	71	65	62	63	58	72

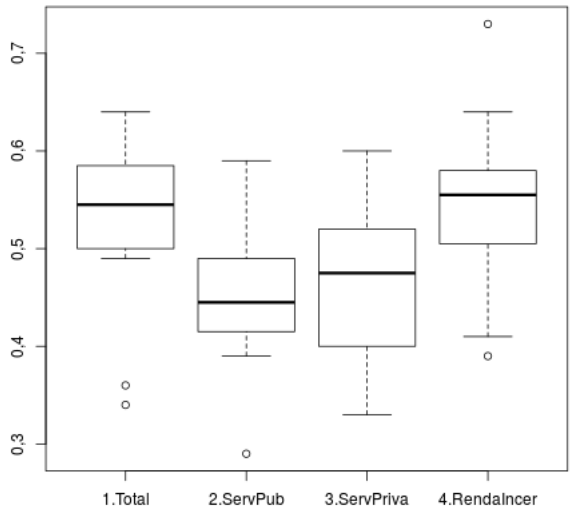

Figura 5.15: Imobiliário I - Modelagem Mensal - Especificidade

A Tabela 5.13 apresenta os resultados obtidos nas validações dos modelos mensais em cada subsegmento de Imobiliário I. Assim como na modelagem anual, estes subsegmentos apresentaram distintos entre si e o segmento completo.

A Figura 5.15 apresenta a comparação da especificidade em cada um dos subsegmentos e do segmento total. Os resultados foram diferentes dos observados na modelagem anual, uma vez que foi possível observar que o subsegmento Setor Público não apresentou uma área de coincidência, considerando os 2º e 3º quartis, com o segmento completo.

A Figura 5.16 apresenta a comparação do indicador Precisão entre os subsegmentos e é possível observar a coincidência de resultados entre os subsegmentos e o segmento completo.

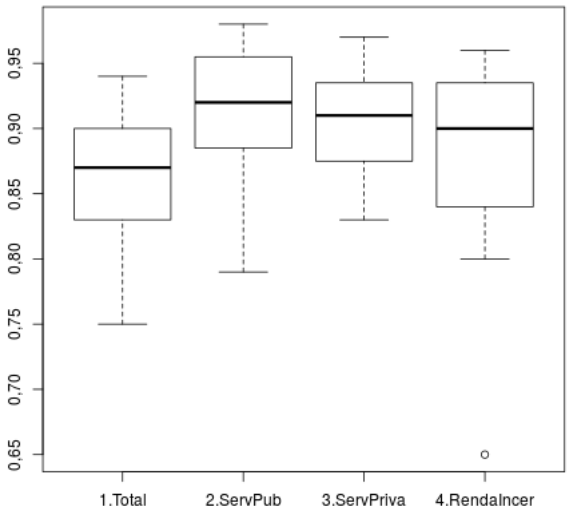

Figura 5.16: Imobiliário I - Modelagem Mensal - Precisão

Finalmente, a Figura 5.8 apresenta o comparativo do indicador Acurácia diante dos subsegmentos e o segmento total de Negócios Sociais, evidenciando a existência de resultados coincidentes entre eles.

Escolha entre Segmento ou Subsegmentos

Foi utilizado o mesmo critério de decisão da análise da modelagem anual, optando-se pela escolha do modelo treinado apenas pelo segmento completo, em razão da existência, na maior parte dos casos, de resultados coincidentes. Esta escolha privilegia o modelo mais generalista, buscando assim, evitar um modelo superajustado.

5.3.3 Comparativo da Modelagem Anual e Mensal

Como esclarecido na análise do segmento Negócios Sociais, o indicador escolhido para a definição do melhor modelo foi a acurácia. A Figura 5.18 compara os resultados da acurácia da modelagem anual e mensal, permitindo-nos observar a grande coincidência de resultados entre as duas modelagens. Mantendo o critério de escolha de um modelo mais generalista, a modelagem anual foi escolhida para ser utilizada no segmento Imobiliário I.

A modelagem, validação, comparação e escolha dos modelos obedeceram aos mesmos critérios apresentados nos segmentos Negócios Sociais e Imobiliário I, que foram detalhados neste documento. Para os próximos segmentos serão apresentados apenas um resumo

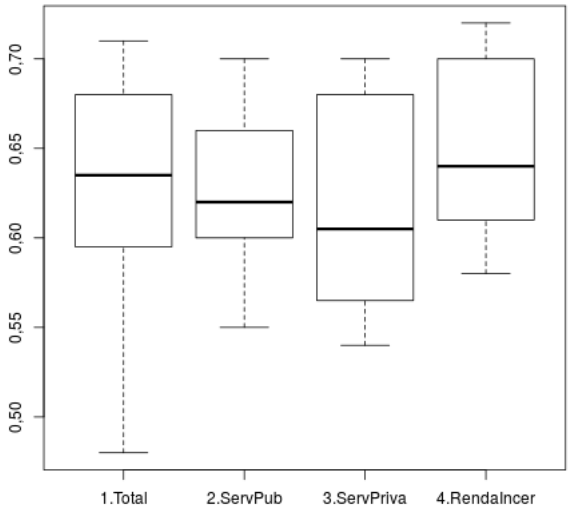

Figura 5.17: Imobiliário I - Modelagem Anual - Acurácia

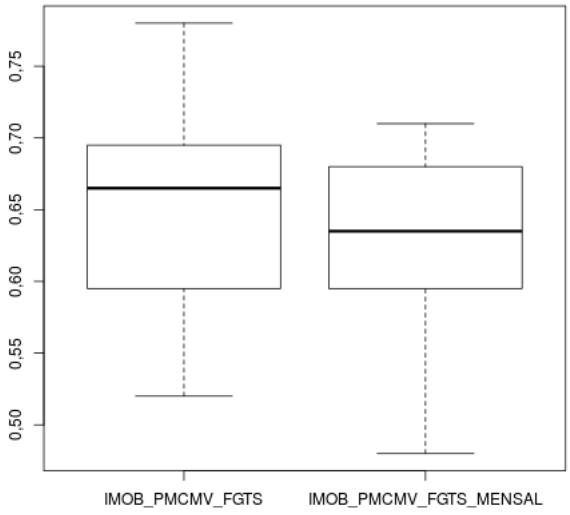

Figura 5.18: Imobiliário I - Comparativo Modelagem Anual x Mensal

Tabela 5.14: Imobiliário II - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Imobiliário II	Total	GBM	ntrees = 400, max_depth = 7
Imobiliário II	Público	GBM	ntrees = 600, max_depth = 7
Imobiliário II	Privado	GLM	alpha = 0.4, lambda = 1e-05
Imobiliário II	Renda Incerta	GLM	alpha = 0.5, lambda = 1e-05

Tabela 5.15: Validações do Segmento Imobiliário II - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	57	57	48	30	50	38	66	54	47	44	56	64
Precisão	76	80	95	98	98	98	79	93	93	99	93	87
Acurácia	63	65	76	47	80	73	69	67	61	62	70	69

dos procedimentos e a decisão dos modelos escolhidos, evitando-se repetições desnecessárias.

Para os próximos segmentos, serão apresentados apenas os resultados da modelagem anual, que foi a estratégia que apresentou os melhores indicadores, à exceção de Negócios Sociais, que já foi apresentada na seção 5.2.

5.4 Imobiliário II

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Imobiliário II e os seus subsegmentos.

A Tabela 5.14 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.15 apresenta a apuração da validação realizada no segmento Imobiliário II. A especificidade variou entre 30% e 66%. Já a precisão apresentou uma variação entre 76% e 99%. A acurácia variou de 47% a 80%.

As Figuras 5.19, 5.20 e 5.21 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Imobiliário II. É possível observar a coincidência de resultados entre os subsegmentos e o segmento total, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

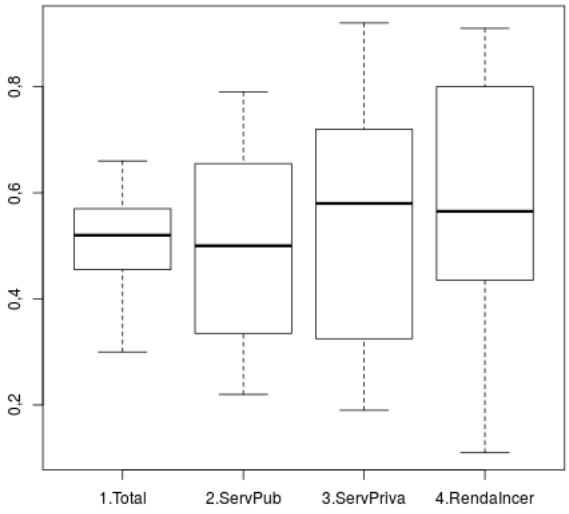

Figura 5.19: Imobiliário II - Modelagem Anual - Especificidade

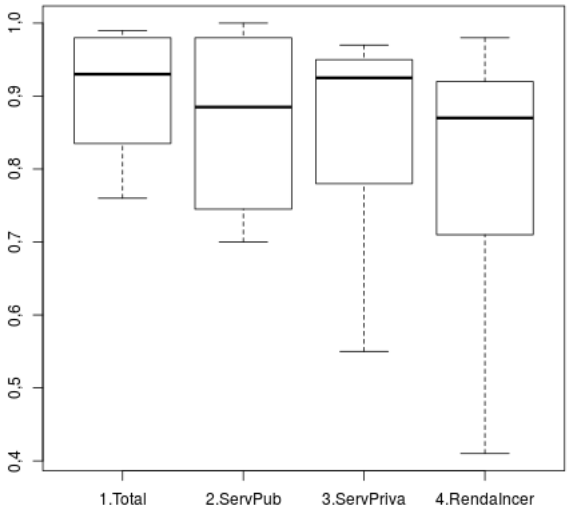

Figura 5.20: Imobiliário II - Modelagem Anual - Precisão

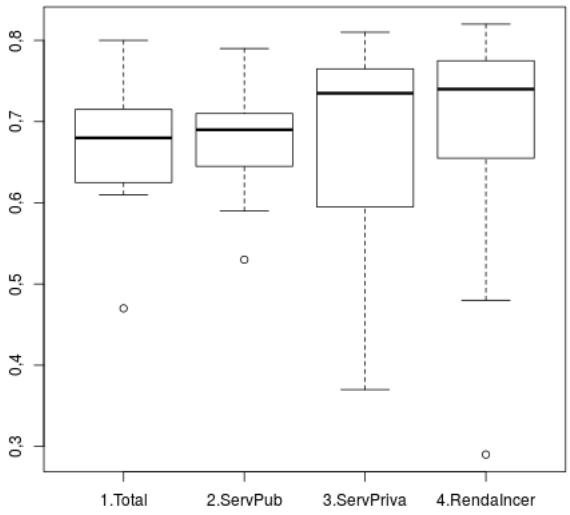

Figura 5.21: Imobiliário II - Modelagem Anual - Acurácia

Tabela 5.16: Imobiliário III - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Imobiliário III	Total	GBM	ntrees = 200, max_depth = 7
Imobiliário III	Público	GBM	ntrees = 200, max_depth = 12
Imobiliário III	Privado	DRF	ntrees = 200, max_depth = 10
Imobiliário III	Renda Incerta	GBM	ntrees = 200, max_depth = 7

5.5 Imobiliário III

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Imobiliário III e os seus subsegmentos.

A Tabela 5.16 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.17 apresenta a apuração da validação realizada no segmento Imobiliário III. A especificidade variou entre 51% e 77%. Já a precisão apresentou uma variação entre 76% e 94%. A acurácia variou de 60% a 79%.

Tabela 5.17: Validações do Segmento Imobiliário III - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	51	61	73	56	67	70	74	77	72	71	51	72
Precisão	88	86	89	91	91	94	90	76	82	81	87	79
Acurácia	60	68	79	65	79	79	78	77	74	74	64	73

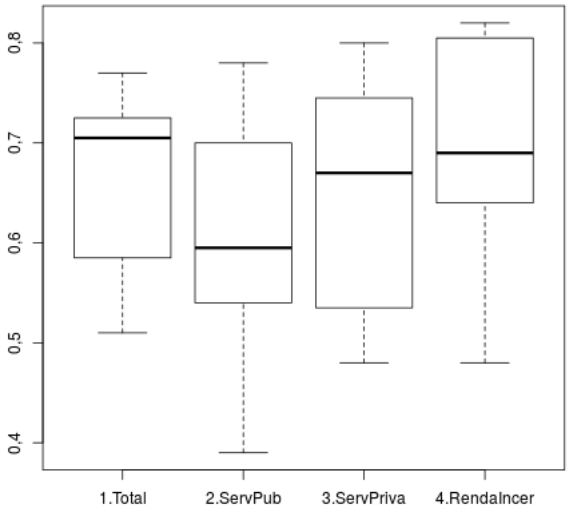

Figura 5.22: Imobiliário III - Modelagem Anual - Especificidade

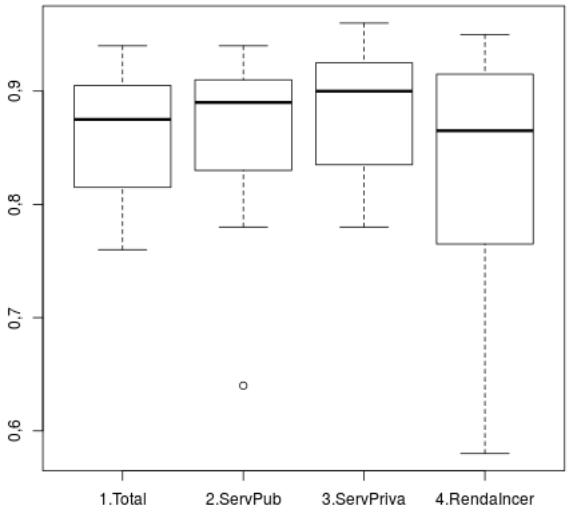

Figura 5.23: Imobiliário III - Modelagem Anual - Precisão

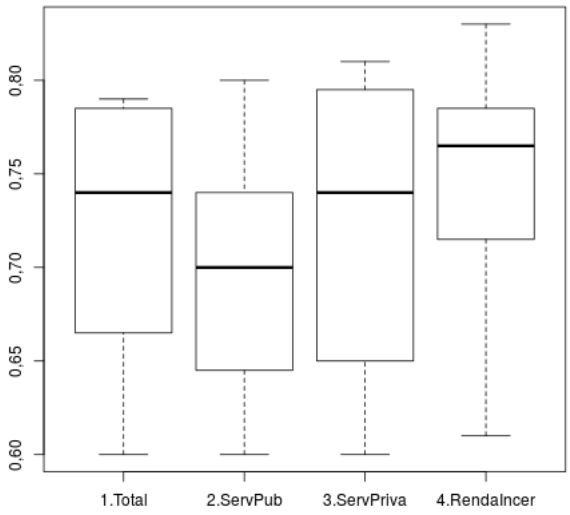

Figura 5.24: Imobiliário III - Modelagem Anual - Acurácia

Tabela 5.18: Veículos I - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Veículos I	Total	GBM	ntrees = 200, max_depth = 10
Veículos I	Público	DL	hidden = (150, 150, 50), epochs = 20
Veículos I	Privado	DRF	ntrees = 200, max_depth = 13
Veículos I	Renda Incerta	DRF	ntrees = 600, max_depth = 13

As Figuras 5.22, 5.23 e 5.24 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Imobiliário III. Assim como ocorreu com os segmentos Imobiliário I e II, os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

5.6 Veículos I

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Veículos I e os seus subsegmentos.

A Tabela 5.18 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.19 apresenta a apuração da validação realizada no segmento Veículos I, dividida pelos seus subsegmentos. No subsegmento Setor Público, a especificidade variou entre 36% e 72%. Já a precisão apresentou uma variação entre 75% e 94%. A acurácia

Tabela 5.19: Validações dos Subsegmentos de Veículos I - Anual

		2016											
	Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Público	Especificidade	55	47	69	60	44	62	64	53	58	62	36	72
	Precisão	82	92	89	78	94	87	88	92	82	82	93	75
	Acurácia	62	62	78	65	68	73	69	62	63	66	54	73
Privado	Especificidade	66	66	86	77	81	81	85	85	80	79	88	82
	Precisão	84	89	79	73	87	88	75	75	80	85	62	81
	Acurácia	69	71	84	76	83	83	84	84	80	80	84	82
Renda Incerta	Especificidade	70	69	79	71	78	87	83	83	75	77	78	83
	Precisão	78	83	84	78	88	76	72	68	78	79	74	70
	Acurácia	71	72	81	72	82	84	82	80	75	77	77	82

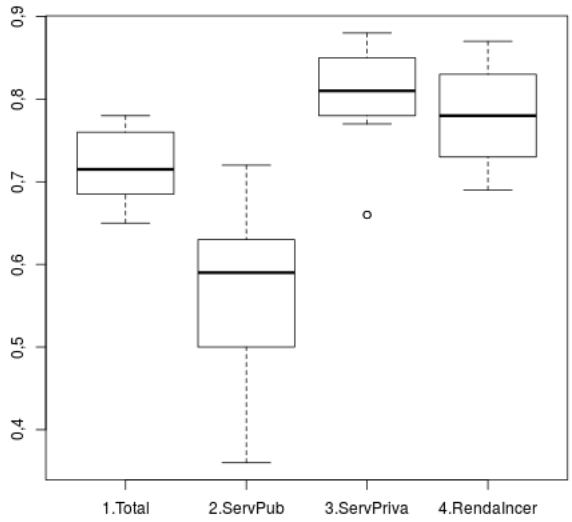

Figura 5.25: Veículos I - Modelagem Anual - Especificidade

variou de 54% a 78%. No subsegmento Setor Privado, a especificidade variou entre 66% e 88%. Já a precisão apresentou uma variação entre 62% e 89%. A acurácia variou de 69% a 84%. No subsegmento Renda Incerta, a especificidade variou entre 69% e 87%. Já a precisão apresentou uma variação entre 68% e 88%. A acurácia variou de 71% a 84%.

As Figuras 5.25, 5.26 e 5.27 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Veículos I. O subsegmento Setor Público evidenciou resultados de Especificidade e Acurácia diferentes dos demais subsegmentos. Nesta situação, optou-se por selecionar a modelagem realizada por subsegmento.

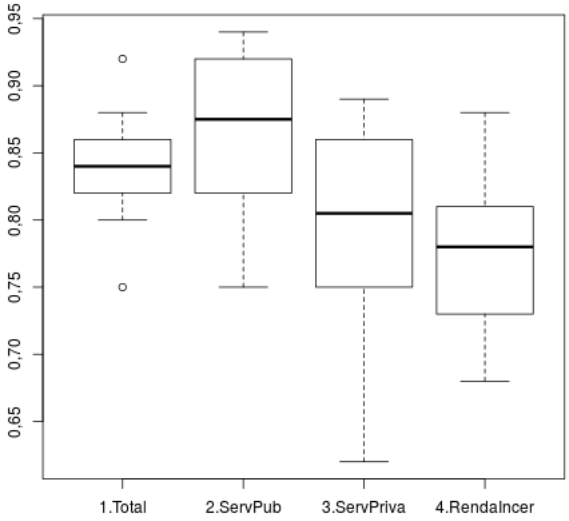

Figura 5.26: Veículos I - Modelagem Anual - Precisão

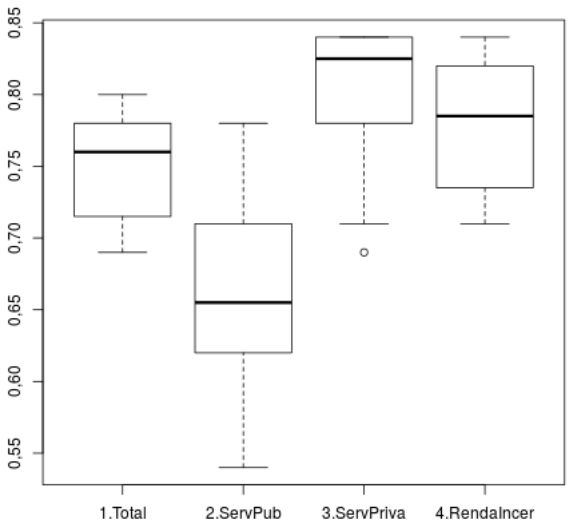

Figura 5.27: Veículos I - Modelagem Anual - Acurácia

Tabela 5.20: Veículos II - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Veículos II	Total	GBM	ntrees = 600, max_depth = 13
Veículos II	Público	DRF	ntrees = 600, max_depth = 7
Veículos II	Privado	GBM	ntrees = 400, max_depth = 13
Veículos II	Renda Incerta	DRF	ntrees = 400, max_depth = 13

Tabela 5.21: Validações do Segmento Veículos II - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	67	80	69	70	67	73	79	80	74	77	73	83
Precisão	85	68	95	82	97	95	87	83	88	85	83	79
Acurácia	70	77	77	72	76	79	80	80	76	78	75	83

5.7 Veículos II

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Veículos II e os seus subsegmentos.

A Tabela 5.20 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.21 apresenta a apuração da validação realizada no segmento Veículos II. A especificidade variou entre 67% e 83%. Já a precisão apresentou uma variação entre 68% e 97%. A acurácia variou de 70% a 83%.

As Figuras 5.28, 5.29 e 5.30 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Veículos II. Os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

5.8 Agronegócios

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Agronegócios e os seus subsegmentos.

A Tabela 5.22 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.23 apresenta a apuração da validação realizada no segmento Agronegócios. A especificidade variou entre 93% e 99%. Já a precisão apresentou uma variação entre 17% e 46%. A acurácia variou de 90% a 98%.

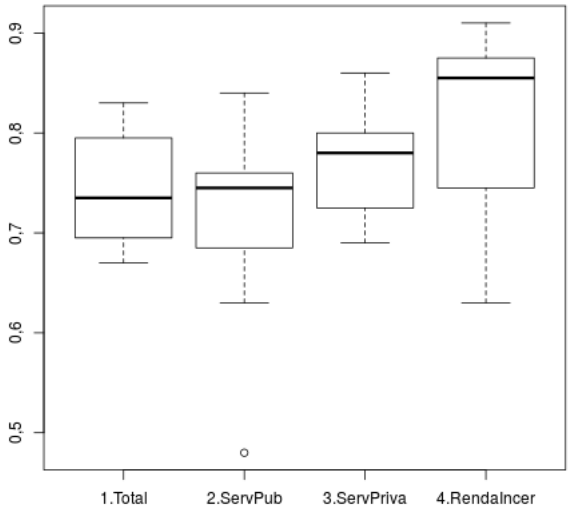

Figura 5.28: Veículos II - Modelagem Anual - Especificidade

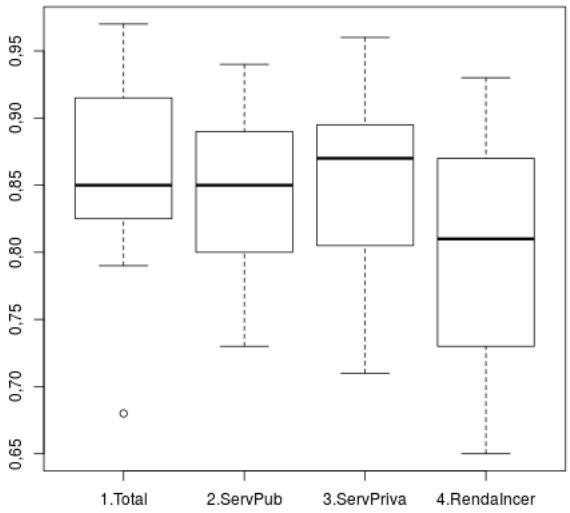

Figura 5.29: Veículos II - Modelagem Anual - Precisão

Tabela 5.22: Agronegócios - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Agronegócios	Total	GBM	ntrees = 400, max_depth = 10
Agronegócios	Público	GBM	ntrees = 200, max_depth = 10
Agronegócios	Privado	GBM	ntrees = 400, max_depth = 10
Agronegócios	Renda Incerta	GBM	ntrees = 600, max_depth = 10

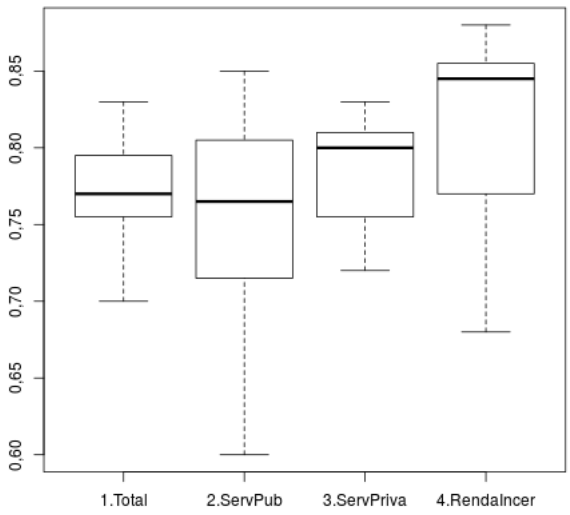

Figura 5.30: Veículos II - Modelagem Anual - Acurácia

Tabela 5.23: Validações do Segmento Agronegócios - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	98	97	97	98	96	99	99	99	99	93	99	98
Precisão	42	42	40	28	46	30	37	25	33	29	17	22
Acurácia	96	95	95	96	94	97	98	98	98	90	98	97

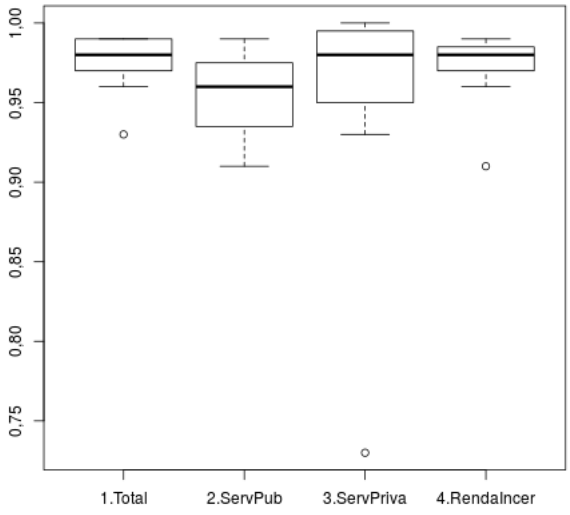

Figura 5.31: Agronegócios - Modelagem Anual - Especificidade

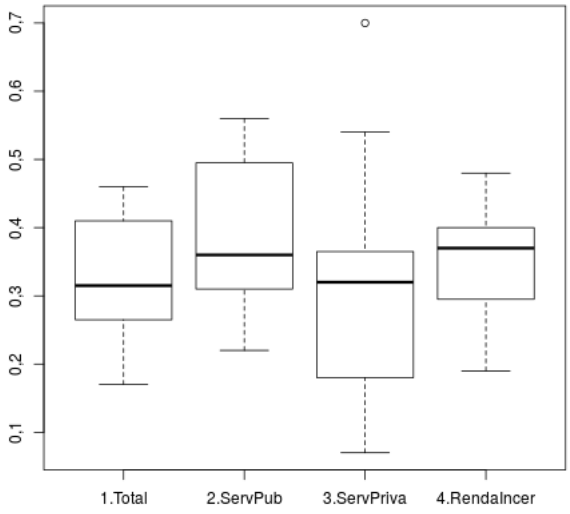

Figura 5.32: Agronegócios - Modelagem Anual - Precisão

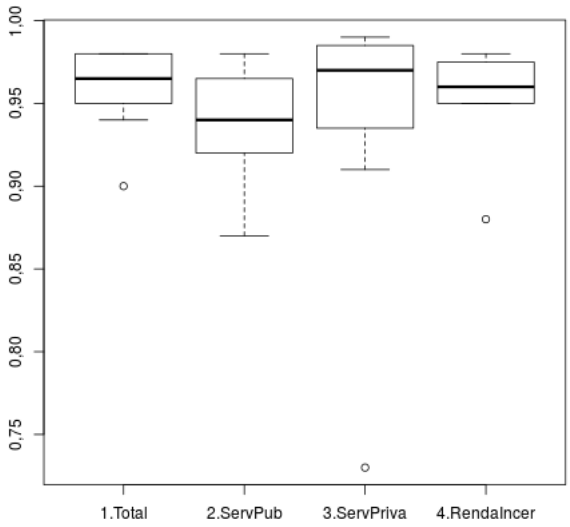

Figura 5.33: Agronegócios - Modelagem Anual - Acurácia

Tabela 5.24: Cartão de Crédito I - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Cartão de Crédito I	Total	DL	hidden = (100, 100, 50), epochs = 25
Cartão de Crédito I	Público	GLM	alpha = 0.6, lambda = 1e-05
Cartão de Crédito I	Privado	GLM	alpha = 0.6, lambda = 1e-03
Cartão de Crédito I	Renda Incerta	DL	hidden = (150, 150), epochs = 20

Tabela 5.25: Validações do Segmento Cartão de Crédito I - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	92	56	91	67	95	94	99	99	99	99	97	99
Precisão	46	91	85	86	88	85	40	42	43	50	56	41
Acurácia	92	56	90	68	93	92	98	98	99	98	96	99

As Figuras 5.31, 5.32 e 5.33 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Agronegócios. Os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

5.9 Cartão de Crédito I

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Cartão de Crédito I e os seus subsegmentos.

A Tabela 5.24 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.25 apresenta a apuração da validação realizada no segmento Cartão de Crédito I. A especificidade variou entre 56% e 99%. Já a precisão apresentou uma variação entre 40% e 91%. A acurácia variou de 56% a 99%.

As Figuras 5.34, 5.35 e 5.36 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Cartão de Crédito I. Os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

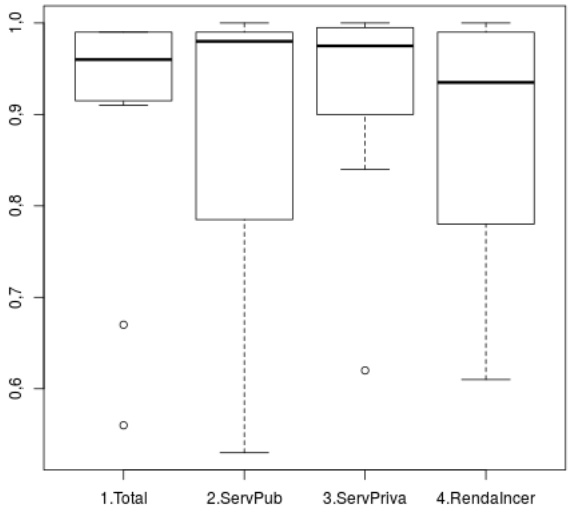

Figura 5.34: Cartão de Crédito I - Modelagem Anual - Especificidade

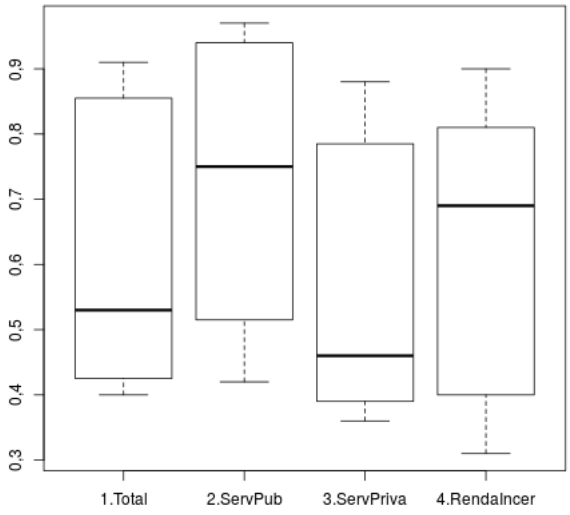

Figura 5.35: Cartão de Crédito I - Modelagem Anual - Precisão

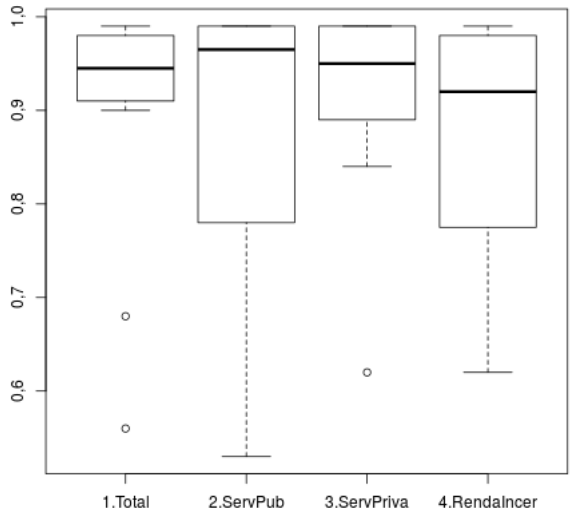

Figura 5.36: Cartão de Crédito I - Modelagem Anual - Acurácia

Tabela 5.26: Cartão de Crédito II - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Cartão de Crédito II	Total	GBM	ntrees = 400, max_depth = 10
Cartão de Crédito II	Público	DRF	ntrees = 200, max_depth = 13
Cartão de Crédito II	Privado	DRF	ntrees = 400, max_depth = 10
Cartão de Crédito II	Renda Incerta	DL	hidden = (50, 50, 50), epochs = 10

5.10 Cartão de Crédito II

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Cartão de Crédito II e os seus subsegmentos.

A Tabela 5.26 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.27 apresenta a apuração da validação realizada no segmento Cartão de Crédito II. A especificidade variou entre 72% e 100%. Já a precisão apresentou uma variação entre 22% e 95%. A acurácia variou de 72% a 99%.

Tabela 5.27: Validações do Segmento Cartão de Crédito II - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	78	72	96	80	96	96	99	99	99	97	97	100
Precisão	90	95	75	75	81	79	51	53	50	51	63	22
Acurácia	78	72	94	80	94	94	98	98	98	96	96	99

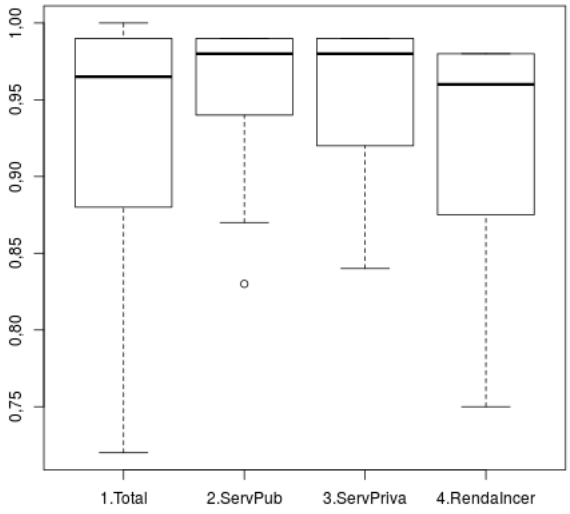

Figura 5.37: Cartão de Crédito II - Modelagem Anual - Especificidade

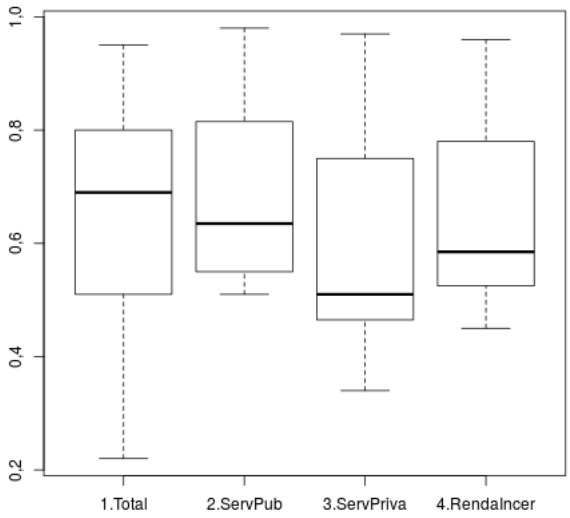

Figura 5.38: Cartão de Crédito II - Modelagem Anual - Precisão

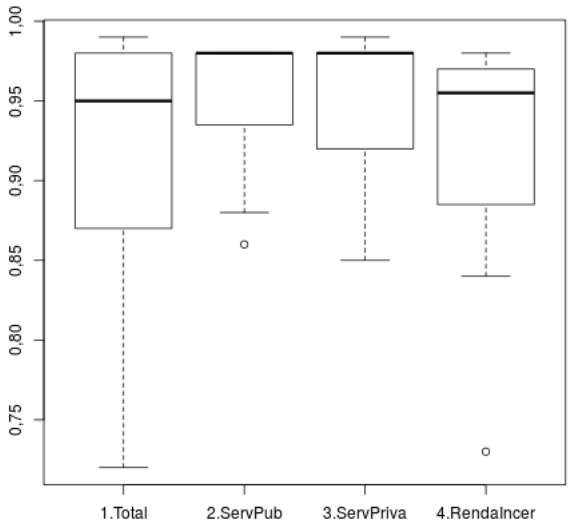

Figura 5.39: Cartão de Crédito II - Modelagem Anual - Acurácia

Tabela 5.28: Demais Operações I - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Demais Operações I	Total	GBM	ntrees = 400, max_depth = 13
Demais Operações I	Público	GBM	ntrees = 200, max_depth = 13
Demais Operações I	Privado	DRF	ntrees = 400, max_depth = 13
Demais Operações I	Renda Incerta	GBM	ntrees = 600, max_depth = 13

As Figuras 5.37, 5.38 e 5.39 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Cartão de Crédito II. Os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

5.11 Demais Operações I

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Demais Operações I e os seus subsegmentos.

A Tabela 5.28 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.29 apresenta a apuração da validação realizada no segmento Demais Operações I. A especificidade variou entre 89% e 99%. Já a precisão apresentou uma variação entre 33% e 84%. A acurácia variou de 88% a 97%.

Tabela 5.29: Validações do Segmento Demais Operações I - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	89	91	96	90	93	96	99	98	97	96	99	98
Precisão	79	74	84	81	84	82	48	54	57	53	33	42
Acurácia	88	89	94	89	92	94	97	95	95	94	95	96

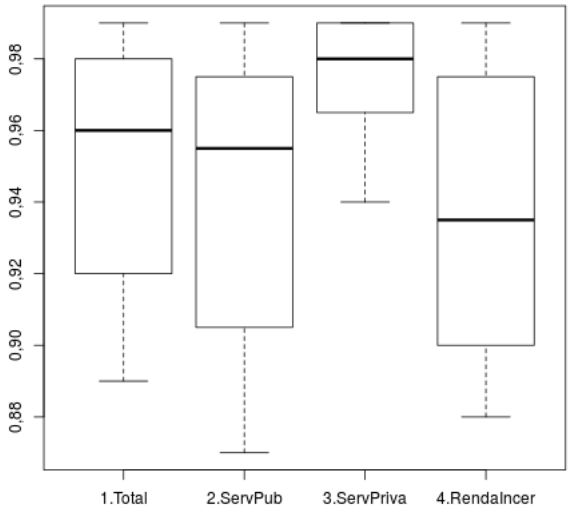

Figura 5.40: Demais Operações I - Modelagem Anual - Especificidade

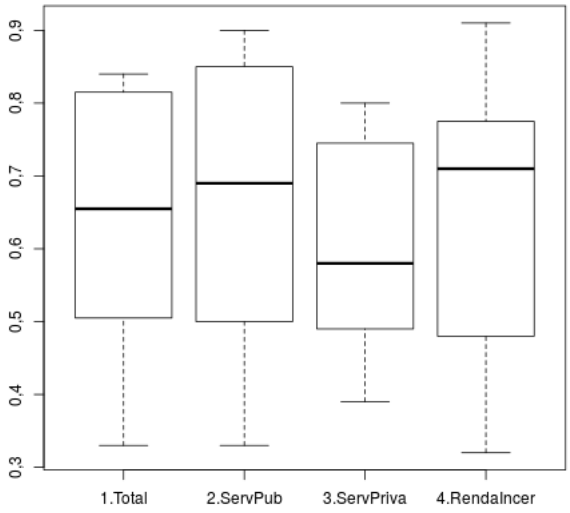

Figura 5.41: Demais Operações I - Modelagem Anual - Precisão

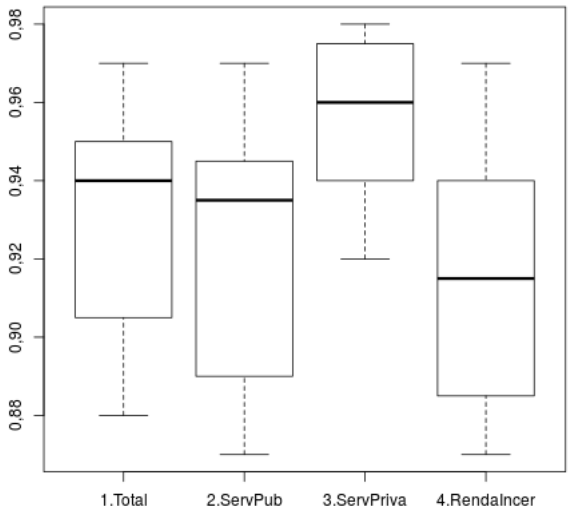

Figura 5.42: Demais Operações I - Modelagem Anual - Acurácia

Tabela 5.30: Demais Operações II - Parâmetros dos Algoritmos

Segmento	Subsegmento	Algoritmo	Parâmetros
Demais Operações II	Total	GBM	ntrees = 400, max_depth = 13
Demais Operações II	Público	DRF	ntrees = 200, max_depth = 13
Demais Operações II	Privado	GBM	ntrees = 400, max_depth = 10
Demais Operações II	Renda Incerta	GBM	ntrees = 200, max_depth = 13

As Figuras 5.40, 5.41 e 5.42 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Demais Operações I. Os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

5.12 Demais Operações II

Esta seção apresentará os resultados obtidos na modelagem, validação e seleção de modelos para o segmento Demais Operações II e os seus subsegmentos.

A Tabela 5.30 apresenta os parâmetros e o algoritmo utilizados na modelagem desse segmento e seus subsegmentos.

A Tabela 5.31 apresenta a apuração da validação realizada no segmento Demais Operações II. A especificidade variou entre 81% e 97%. Já a precisão apresentou uma variação entre 42% e 94%. A acurácia variou de 82% a 94%.

Tabela 5.31: Validações do Segmento Demais Operações II - Anual

	2016											
Indicador(%)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Especificidade	84	87	81	90	86	92	95	94	93	94	97	96
Precisão	82	67	89	61	94	82	68	67	72	65	42	57
Acurácia	84	84	82	86	87	90	93	92	92	92	93	94

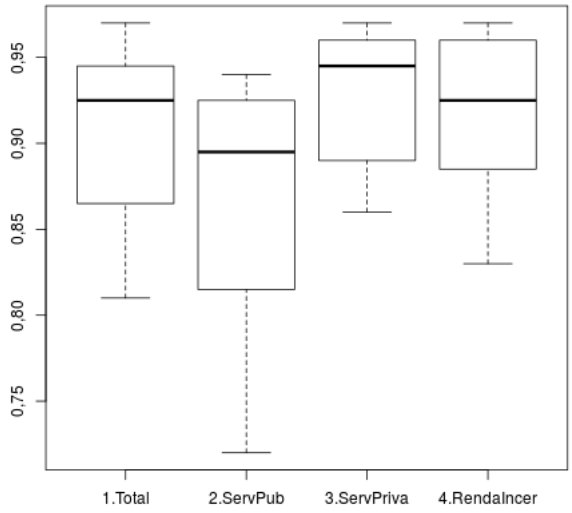

Figura 5.43: Demais Operações II - Modelagem Anual - Especificidade

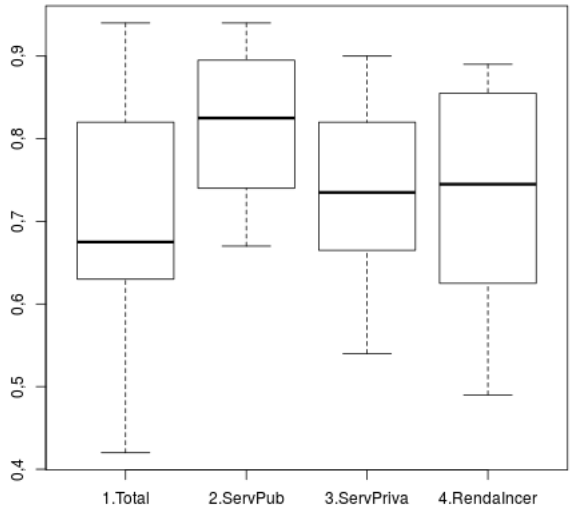

Figura 5.44: Demais Operações II - Modelagem Anual - Precisão

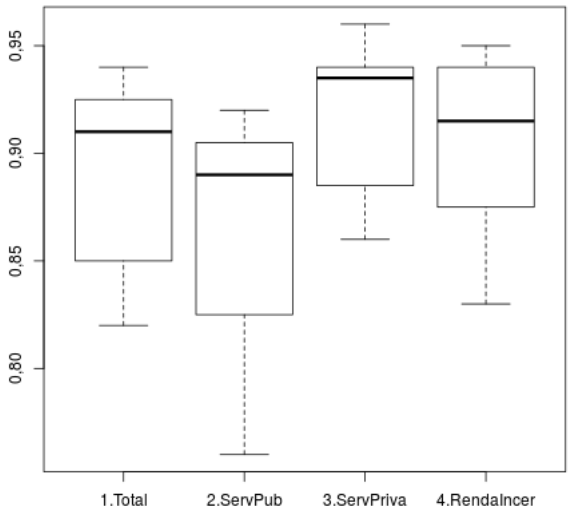

Figura 5.45: Demais Operações II - Modelagem Anual - Acurácia

As Figuras 5.43, 5.44 e 5.45 apresentam os *boxplots* da especificidade, precisão e acurácia, respectivamente, da modelagem anual para os subsegmentos e o segmento total de Demais Operações II. Os resultados entre os subsegmentos e o segmento total apresentaram coincidência de resultados, razão pela qual foi escolhida a modelagem considerando apenas o segmento total.

5.13 Modelos Selecionados

A Tabela 5.32 lista o resumo dos modelos selecionados após a análise das estratégias de segmentação por origem da renda e a perspectiva de treinamento, mensal ou anual, o algoritmo utilizado. A maior parte dos segmentos mostrou-se mais adequada às estratégias de treinamento Anual, sem a necessidade de segmentação por origem da renda. Porém, dois segmentos mostraram-se mais sensíveis a esta segmentação, enquanto apenas um obteve melhores resultados com a modelagem mensal.

Apesar de terem sido utilizados 4 algoritmos de classificação, observou-se que apenas um deles não foi selecionado, o algoritmo *Generalized Linear Models* (GLM). Este algoritmo representa a única técnica utilizada atualmente no Banco Alfa, ou seja, a regressão logística.

Tabela 5.32: Modelos Selecionados

Segmento	Origem da Renda	Treinamento	Algoritmo
Negócios Sociais	Setor Público	Mensal	GBM
	Setor Privado	Mensal	GBM
	Renda Incerta	Mensal	GBM
Imobiliário I	Todos	Anual	DL
Imobiliário II	Todos	Anual	GBM
Imobiliário III	Todos	Anual	GBM
Veículos I	Setor Público	Anual	DL
	Setor Privado	Anual	DRF
	Renda Incerta	Anual	DRF
Veículos II	Todos	Anual	GBM
Agronegócios	Todos	Anual	GBM
Cartões de Crédito I	Todos	Anual	DL
Cartões de Crédito II	Todos	Anual	GBM
Demais Operações I	Todos	Anual	GBM
Demais Operações II	Todos	Anual	GBM

Capítulo 6

Análise dos Resultados

Este capítulo é destinado à análise dos resultados obtidos com os modelos preditivos gerados para os 11 segmentos de operações de crédito do Banco Alfa.

O objetivo desta análise é levantar as hipóteses de utilização dos modelos, verificar sua aplicabilidade e estimar os seus potenciais benefícios.

6.1 Estratégia Atual de Recuperação

Conforme explicado anteriormente, a estratégia de recuperação de crédito do Banco Alfa é realizada conforme o segmento que a operação em atraso se enquadra. Esses segmentos são caracterizados pelo tipo de operação de crédito e, em alguns casos, o nível de renda do cliente responsável pela operação.

Há uma processo sistemático de abordagem desses clientes inadimplentes, que envolve o envio de mensagens de texto, cartas de aviso, ligações telefônicas realizadas por funcionários e por empresa especializada de cobrança. Esse processo é iniciado após a observação de 15 dias de atraso de uma operação, ou seja, justamente o prazo mínimo de atraso das operações estudadas nesta pesquisa.

Este estudo não obteve acesso aos dados que permitiriam a identificação das abordagens realizadas aos clientes inadimplentes. Considerando que esta abordagem está regulamentada como um procedimento obrigatório na instituição, para fins desta análise, tomaremos por pressuposto que todos os clientes foram abordados.

6.2 Estratégia Proposta de Utilização dos Modelos

Antes de apresentar uma nova estratégia de recuperação de crédito, faz-se necessário detalhar alguns conceitos, que direcionaram a elaboração dessa proposta:

- Os modelos preditivos foram treinados para classificar as operações em duas classes: com potencial de recuperação e sem potencial de recuperação. A geração desses rótulos foi condicionada à ocorrência de recuperação em um prazo máximo de 30 dias, ou seja, quando afirma-se ter potencial de recuperação, está sendo predito que haverá recuperação da inadimplência no prazo máximo de 30 dias. Por outro lado, ao se afirmar que não há potencial de recuperação, não está sendo predito que não haverá recuperação, mas apenas que isto não ocorrerá nos próximos 30 dias.

- A análise descritiva apontou que a variável Número de Dias em Atraso possui uma correlação negativa com a variável resposta, ou seja, quanto maior for o atraso, menores a chances de recuperação.

- À medida que há um aumento no atraso da operação eleva-se a PCLD, que pode atingir até 100% do valor da operação.

- Em 03 de abril de 2017 entrou em vigor a Resolução 4.596 do BACEN, alterando as regras de utilização do crédito rotativo do cartão de crédito, limitando o seu uso por no máximo 30 dias. Após esse prazo, saldo da fatura deve ser convertido para uma modalidade de crédito mais vantajosa ao devedor. Esta nova regra provoca um grande impacto no perfil das operações de cartão de crédito em atraso, invalidando a utilização dos modelos produzidos neste estudo.

- Cada segmento de operações possui uma estratégia personalizada de recuperação e envolve diversas ações, que são iniciadas à medida que a operação atinge uma determinada quantidade de dias em atraso. Como exemplo hipotético dessas ações, onde x representa um número inteiro positivo, pode-se citar:

  - 15 dias de atraso - bloqueia renovação do cheque especial.
  - 15 + x dias de atraso - cancelamento de limite de crédito.
  - 15 + 2x dias de atraso - envio ao SPC/Serasa.
  - 15 + 3x dias de atraso - cancela cheque especial.
  - 15 + 10x dias de atraso - indicação de cessão da dívida,

- O canal utilizado para abordagem ao cliente inadimplente também é estabelecido conforme a quantidade de dias em atraso. À medida em que há um aumento neste atraso, promove-se a mudança da canal utilizado. Um exemplo hipotético dessa mudança, onde x representa um número inteiro positivo, pode-se citar:

  - 15 + x dias de atraso - agência

- 15 + 2x dias de atraso - central de atendimento
- 15 + 3x dias de atraso - empresa terceirizada

Com base nessas considerações, é possível adotar as seguintes estratégias de recuperação de crédito, conforme a classificação predita pelos modelos:

- Sem potencial de recuperação - Presume-se que as operações enquadradas nesta classe promoverão um aumento de PCLD no mês subsequente. Além disto, as possibilidades de recuperação serão ainda menores no mês seguinte, caso não haja mudança nas características do cliente. Portanto, propõe-se a diminuição em 30 dias dos prazos estabelecidos para as ações previstas para o segmento da operação de crédito, antecipando-se a execução da estratégia de cobrança e a mudança do canal de abordagem ao cliente.

- Com potencial de recuperação - Para essa classe, propõe-se a intensificação da abordagem ao cliente, mantendo o mesmo canal de abordagem pelos próximos 30 dias, evitando-se a alteração do interlocutor do Banco com o cliente. A abordagem seria realizada para todas as operações enquadradas nesta classe, todavia seguindo uma prioridade conforme a pontuação calculada pelos modelos, iniciando a abordagem pelas operações com a maior pontuação.

6.3 Avaliação do Impacto da Estratégia Sugerida

Para avaliar o impacto gerado pela estratégia sugerida por esse estudo, foi realizada uma classificação das operações com uma base de testes elaborada com todas as operações de crédito em atraso superior a 14 dias existentes em 31 de janeiro de 2017. Essa classificação foi confrontada com a observação dessas mesmas operações em 28 de fevereiro de 2017, que permitiu aferir a performance dessa classificação. Estas operações não foram utilizadas no treinamento, nem mesmo da validação dos modelos produzidos neste estudo. Podem ser classificados como dados novos.

A classificação realizada no segmento Negócios Sociais apontou 138.952 contratos sem potencial de recuperação e destes, 135.454 contratos efetivamente não apresentaram recuperação após 30 dias. Por outro lado, 27.387 contratos foram classificados com potencial de recuperação e 17.649 apresentaram diminuição no atraso ou diminuição do saldo devedor em 30 dias. A especifidade observada foi de 93%, a precisão 64% e a acurácia 92%.

Em uma visão de PCLD, as operações classificadas como sem recuperação representaram R$ 105,90 milhões, sendo que R$ 105,40 milhões efetivamente não foram recuperados,

Tabela 6.1: Resultados da Simulação da Antecipação de Terceirização de Cobrança

	Sem potencial de recuperação - Terceirização de Cobrança					
Segmento	Quantidade	Acerto		PCLD	Acerto	
		(%)	Qtde	(R$)	(%)	R($)
Negócios Sociais	12.456	91,4	11.388	1.150.000	91,0%	1.046.000
Imobiliário I	9.556	95,5	9.126	92.700.000	98,8%	91.600.000
Imobiliário II	45	100	45	33.500	100,0%	33.500
Imobiliário III	744	95,8	713	17.750.000	99,7%	17.690.000
Veículos I	2.590	82,3	2.132	3.880.000	86,1%	3.340.000
Veículos II	1.900	91,8	1.745	1.490.000	91,9%	1.370.000
Agronegócio	15.876	97,5	15.481	55.330.000	99,6%	55.110.000
Demais Operações I	24.257	90,2	21.879	43.770.000	90,7%	39.710.000
Demais Operações II	150.226	90,0%	135.145	36.590.000	91,6%	33.510.000

ou seja, 99,6% do total predito como sem potencial de recuperação. As operações classificadas com potencial de recuperação representaram R$ 2,5 milhões, sendo que R$ 1,06 milhões foram efetivamente recuperados, ou seja, 41,4% do total predito com potencial de recuperação.

Caso a estratégia proposta por este estudo fosse adotada, os seguintes resultados seriam observados:

- Sem potencial de recuperação - Antecipação de ações:
  - Conforme ilustrado na Tabela 6.1, 12.456 operações seriam encaminhadas para empresa terceirizada de cobrança, representando R$ 1.1 milhão em PCLD. Deste total, 91.4% das operações confirmaram-se sem recuperação, representando R$ 1.04 milhão.
  - A Tabela 6.2 mostra que 3.871 operações teriam a indicação para cessão (venda) da dívida antecipada, representando R$ 3.5 milhões em PCLD. Deste total, 100% das operações confirmaram-se sem recuperação.
- Com potencial de recuperação - Intensificação na abordagem:
  - Conforme listado na Tabela 6.3, 1.220 operações teriam a alteração de canal de abordagem congelados por 30 dias, mantendo-se o interlocutor, representando R$ 175 mil de PCLD. Deste total, 59% das operações confirmaram-se com recuperação, representando R$ 78 mil de PCLD.

Tabela 6.2: Resultados da Simulação da Antecipação de Cessão de Dívida

	Sem potencial de recuperação - Antecipação de Cessão de Dívida					
Segmento	Quantidade	Acerto		PCLD	Acerto	
		(%)	Qtde	(R$)	(%)	R($)
Negócios Sociais	3.871	100	3.871	3.500.000	100	3.500.000
Imobiliário I	740	97,8	724	44.350.000	98,1	43.504.000
Imobiliário II	3	100	3	49.000	100	49.000
Imobiliário III	127	100	127	16.800.000	100	16.800.000
Veículos I	426	100	426	15.000.000	100	15.000.000
Veículos II	568	100	568	6.900.000	100	6.900.000
Agronegócio	676	100	676	16.000.000	100	16.000.000
Demais Operações I	7.940	100	7.940	105.000.000	100	105.000.000
Demais Operações II	88.870	99,9	88.802	197.000.000	99,9	196.900.000

Tabela 6.3: Resultados da Simulação da Manutenção do Canal de Cobrança

	Com potencial de recuperação - Manutenção do canal de cobrança					
Segmento	Quantidade	Acerto		PCLD	Acerto	
		(%)	Qtde	(R$)	(%)	R($)
Negócios Sociais	1.220	59,1	721	175.000	44,6	78.000
Imobiliário I	10.930	36,7	4.009	102.441.000	31,6	32.378.000
Imobiliário II	105	41,0	43	85.000	41,2	3.000
Imobiliário III	706	40,8	288	9.387.000	15,9	1.489.000
Veículos I	2.342	36,5	854	4.598.000	25,9	1.193.000
Veículos II	2.055	45,6	937	1.799.000	37,4	672.000
Agronegócio	16.033	3,8	610	59.522.000	3,7	2.192.000
Demais Operações I	1.755	39,0	684	2.967.000	29,9	886.000
Demais Operações II	43.131	35,3	15.207	15.014.000	19,3	2.891.000

6.4 Benefícios Esperados

As estratégias propostas podem aprimorar o controle da PCLD do Banco Alfa com os seguintes benefícios:

- Antecipação da Terceirização de Cobrança - Esta estratégia pode desonerar as atividades das agências, evitando o dispêndio de esforços dos funcionários nas atividades de cobrança, uma vez que a predição desses casos apresentou uma alta taxa de acerto, variando de 82.3% a 100%.

- Antecipação da Indicação de Cessão da Dívida - Há dois benefícios esperados na utilização desta estratégia. O primeiro deles é a redução dos custos com a realização das cobranças, uma vez que a dívida será cedida. O segundo está no fato de que, ao antecipar esta cessão, esses ativos podem ser vendidos com um valor maior do que se fossem negociados 30 dias mais tarde, uma vez que apresentarão um número menor de dias em atraso.

O próximo capítulo apresentará as discussões finais deste trabalho, com sua conclusão e sugestão de trabalhos futuros.

Capítulo 7

Conclusões e Trabalhos Futuros

Este capítulo aborda as conclusões, os resultados obtidos, e apresenta os trabalhos futuros que podem ser conduzidos a partir das realizações alcançadas com este trabalho.

7.1 Conclusões

Esse trabalho foi realizado com o objetivo de auxiliar o Banco Alfa na redução de sua PCLD, visando identificar os clientes com maior potencial de regularização de suas operações. Em seguida foram estabelecidos alguns objetivos específicos que, se atingidos, promoveriam o alcance do objetivo geral deste estudo.

O objetivo específico 1 deste trabalho foi a elaboração de modelos preditivos conforme o perfil da renda do cliente e compará-los a modelos sem a distinção desse perfil. Para isso, cada um dos 11 segmentos foi dividido em subsegmentos de acordo com o perfil da renda do cliente: setor público, privado ou renda incerta.

A realização desta comparação mostrou-se bem sucedida, uma vez que foram identificados alguns segmentos que se adaptaram melhor a um modelo preditivo que considerava o perfil do cliente. Os segmentos Negócios Sociais e Veículos I apresentaram melhores resultados com esta segmentação. Para os demais modelos, manteve-se a decisão de desconsiderar o perfil de renda do cliente, pois a comparação dos resultados não mostrou a superioridade de uma estratégia frente à outra. Nesta condição, selecionou-se o modelo mais generalizado.

O objetivo específico 2 era elaborar modelos preditivos conforme o tipo de produto consumido e compará-los a um modelo preditivo único, independente do produto. A elaboração deste último tipo de modelo mostrou-se infrutífera, uma vez que todas regras de negociação de contratos atrasados do Banco Alfa eram estabelecidas de acordo com o produto consumido. Faziam parte dessas regras todo o fluxo de negociação e cobrança,

os parâmetros de valores e descontos a serem concedidos e, consequentemente, as alçadas de negociações.

Além destes motivos, os produtos possuíam características básicas tão distintas, como prazo de pagamento e valores contratados, que já na análise descritiva percebeu-se a necessidade de manter os produtos separados, evitando-se a percepção de inúmeros *outliers*. Um exemplo disto, seria incluir em uma mesma base de treinamento operações de crédito imobiliário e financiamento de veículos. Enquanto o primeiro possui um prazo de pagamento geralmente acima dos 240 meses e valores contratados acima de R$ 100 mil, os financiamentos de veículos não ultrapassavam os 48 meses e, em geral, eram abaixo de R$ 100 mil.

Neste contexto, optou-se por dar continuidade à pesquisa levando-se em conta as regras de negociação de operações em atraso do Banco, que resultou na identificação de 11 segmentos

Por sua vez, o objetivo específico 3 era elaborar modelos preditivos para cada mês de observação e compará-los a um modelo preditivo que leve em conta as observações independente do mês. Para viabilizar esta comparação definiu-se uma estratégia treinar os modelos conforme cada segmento, com todos os meses de 2015, comparando-os com modelos treinados em cada mês desse mesmo ano. As bases de validações foram exatamente a mesmas para os dois casos, ou seja, os 12 meses do ano 2016, o que permitiu uma comparação precisa deste caso.

Ainda que a maioria dos segmentos tenha se adaptado melhor à modelagem anual, o segmento Negócios Sociais apresentou melhores resultados utilizando-se a modelagem mensal, aumentando a performance de seus indicadores.

Já o objetivo 4 compreendia em propor a criação de um índice que auxiliasse na priorização da abordagem ao cliente devedor na tentativa de renegociação de suas operações. Este índice foi criado, sendo obtido por meio da escoragem dos modelos de classificação. Todavia, a estratégia definida pelo Banco Alfa determinava a abordagem de todos os clientes, ainda que fosse necessária a contratação de outros canais para realizar esta atividade.

Desta maneira, mudou-se apenas a destinação do índice proposto. Em vez de utilizá-lo na priorização da abordagem, o índice servirá para antecipar as ações de cobrança ou manutenção de canais de abordagem. Para as operações com as menores pontuações, ou seja, sem potencial de recuperação, o indicador servirá para antecipar a indicação de cessão (venda) da dívida ou a antecipação da transferência do canal de negociação para empresas terceirizadas, que já prestam este serviço para o Banco.

Quando se tratar de operações com alta pontuação do indicador, ou seja, com potencial de recuperação, poderá ser proposta a manutenção do canal de negociação, ainda que

as operações atinjam atraso suficiente para que seja alterado esse canal. A precisão observada na utilização desta estratégia não foi alta. Em várias simulações, a quantidade de operações que foram preditas como recuperáveis, não se confirmaram. Todavia, estima-se que todas elas sofreram a mudança de canal de negociação, uma vez que esta estratégia ainda não foi implantada. Espera-se que, ao manter-se o interlocutor com o cliente com potencial de recuperação, aumente-se a quantidade de casos recuperados.

Por último, declarou-se objetivo específico 4, que era calcular a estimativa de montantes financeiros a serem recuperados com a estratégia proposta. Na simulação realizada, utilizando-se dados de janeiro de 2017, apontou-se cerca de R$ 195 milhões em PCLD com potencial de recuperação. Todavia, ainda que todas as operações fossem recuperadas, há outros fatores que determinariam o novo valor da PCLD, como a existência de atraso do cliente em outras operações e o risco de crédito do cliente.

Apesar de não ser possível estimar os valores financeiros recuperáveis, espera-se uma diminuição das despesas operacionais do processo de recuperação de crédito. Por exemplo, com a antecipação da cessão da dívida, é possível que se consiga negociá-la em condições mais favoráveis.

Além da declaração dos objetivos específicos já expostas nos parágrafos anteriores, algumas hipóteses foram formuladas de modo a direcionar o estudo. Ao atingir os objetivos definidos neste trabalho, tornou-se possível avaliar estas hipóteses.

A primeira delas referia-se à expectativa de que o perfil da renda do cliente afetava a inadimplência. Esta hipótese não pode ser refutada, uma vez que os segmentos Negócios Sociais e Veículos I apresentaram-se sensíveis ao perfil da renda. Todavia, nem todos os segmentos foram influenciados.

A segunda hipótese, que referia-se à possibilidade do produto de crédito afetar o comportamento do cliente perante a inadimplência, não pôde ser avaliada. Apesar de ter sido percebido índices diferentes de recuperação das operações em segmentos distintos, não foi possível avaliar se um mesmo cliente se comportaria de forma diferente caso tivesse mais de um tipo de produto em atraso. Todo o estudo foi realizado na perspectiva da operação, sem considerar a avaliação integral do cliente.

Por último, a hipótese de que o comportamento do cliente perante a recuperação da inadimplência pudesse estar ligado a épocas específicas de um ano, também não pôde ser refutada. O comportamento da inadimplência do segmento Negócios Sociais mostrou-se fortemente ligado ao mês em que era feita a tentativa de recuperação. Os demais segmentos não se mostraram afetados por esta característica.

7.2 Resultados Obtidos

A realização deste estudo atingiu alguns resultados e contribuições que passam a ser listados a seguir:

- A identificação de novas estratégias de recuperação de crédito, com a possibilidade de redução de despesas operacionais.
- Foram identificadas características que afetam a inadimplência, como o perfil da renda do cliente e época do ano em que são realizadas as tentativas de negociação, abrindo caminho para novos estudos que explorem estas características.
- A utilização de técnicas de *machine learning* foi utilizada pela primeira vez pelo Banco Alfa para a predição da recuperação de crédito. Com os resultados obtidos, especialmente na identificação dos clientes sem potencial de recuperação, passa-se a ter mais um instrumento para ser utilizado na tomada de decisão.
- A identificação que a única técnica de modelagem preditiva utilizada pelo Banco Alfa atualmente, a regressão logística, apresentou resultados piores comparados às outras três técnicas utilizadas.
- Os estudos iniciais desta pesquisa foram aceitos para apresentação na IEEE - *Internacional Conference on Machine Learning and Applications* (ICMLA) em Los Angeles, CA (USA) em dezembro de 2016 e teve sua publicação nos anais da conferência [15].
- Um versão estendida do artigo publicado em [15] foi aceita e publicada na revista eletrônica *Advances in Science, Technology and Engineering Systems Journal* (ASTESJ)[16].[1]

7.3 Trabalhos Futuros

Algumas das hipóteses que foram assumidas neste trabalho e não puderam ser avaliadas, como o comportamento de um cliente perante tipos diferentes de operações de crédito em atraso. Um novo estudo pode ser realizado sob a ótica integral do cliente, em vez de focar nas operações como feito nesta pesquisa.

Além disto, pode-se realizar novos estudos que considerem as formas de abordagens que foram feitas ao cliente inadimplente, visando avaliar a eficácia delas e utilizar esta informação para a proposição de novas estratégias de recuperação.

[1] Os dois artigos publicados estão

Nesta pesquisa foram estudados apenas os clientes PF, que representavam a maior quantidade de contratos inadimplentes, impactando os processos operacionais de recuperação. Porém, apesar de um número menor de contratos, há um volume financeiro substancialmente maior de inadimplentes no segmento de clientes Pessoa Jurídica. Um estudo para este segmento de clientes também poderá trazer benefícios para o Banco Alfa no controle de sua PCLD.

Referências

[1] *Política Monetária e Operações de Crédito do SFN.* http://www.bcb.gov.br/?ECOIMPOM, acesso em 2016-06-12TZ. 1, 10

[2] Ha, Sung Ho: *Behavioral assessment of recoverable credit of retailer's customers.* Inf. Sci., 180(19):3703–3717, outubro 2010, ISSN 0020-0255. http://dx.doi.org/10.1016/j.ins.2010.06.012. 4, 7

[3] Lessmann, Stefan, Bart Baesens, Hsin Vonn Seow e Lyn C. Thomas: *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.* European Journal of Operational Research, 247(1):124–136, 2015. http://dblp.uni-trier.de/db/journals/eor/eor247.html#LessmannBST15. 4, 7, 8

[4] LeCun, Yann, Yoshua Bengio e Geoffrey Hinton: *Deep learning.* Nature, 521(7553):436–444, 2015. 4, 31

[5] Baesens, Bart, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens e Jan Vanthienen: *Benchmarking state-of-the-art classification algorithms for credit scoring.* Journal of the operational research society, 54(6):627–635, 2003. 7, 31

[6] Verbeke, Wouter, Karel Dejaeger, David Martens, Joon Hur e Bart Baesens: *New insights into churn prediction in the telecommunication sector: A profit driven data mining approach.* European Journal of Operational Research, 218(1):211 – 229, 2012, ISSN 0377-2217. http://www.sciencedirect.com/science/article/pii/S0377221711008599. 8

[7] Dejaeger, Karel, Frank Goethals, Antonio Giangreco, Lapo Mola e Bart Baesens: *Gaining insight into student satisfaction using comprehensible data mining techniques.* European Journal of Operational Research, 218(2):548 – 562, 2012, ISSN 0377-2217. http://www.sciencedirect.com/science/article/pii/S0377221711010137. 8

[8] Caruana, Rich, Art Munson e Alexandru Niculescu-Mizil: *Getting the most out of ensemble selection.* Em *Data Mining, 2006. ICDM'06. Sixth International Conference on*, páginas 828–833. IEEE, 2006. 9

[9] Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer e Rudiger Wirth: *Crisp-dm 1.0 step-by-step data mining guide.* Relatório Técnico, The CRISP-DM consortium, August 2000. http://www.crisp-dm.org/CRISPWP-0800.pdf. 10

[10] Arndt, Stephan, Carolyn Turvey e Nancy C Andreasen: *Correlating and predicting psychiatric symptom ratings: Spearmans r versus kendalls tau correlation*. Journal of psychiatric research, 33(2):97–104, 1999. 16

[11] Japkowicz, Nathalie e Shaju Stephen: *The class imbalance problem: A systematic study*. Intelligent Data Analysis, páginas 429–449, 2002. 32

[12] Brown, Iain e Christophe Mues: *An experimental comparison of classification algorithms for imbalanced credit scoring data sets*. Expert Systems with Applications, 39(3):3446 – 3453, 2012, ISSN 0957-4174. http://www.sciencedirect.com/science/article/pii/S095741741101342X. 32

[13] Wood, Simon N: *Stable and efficient multiple smoothing parameter estimation for generalized additive models*. Journal of the American Statistical Association, 99(467):673–686, 2004. http://dx.doi.org/10.1198/016214504000000980. 32

[14] McGill, Robert, John W. Tukey e Wayne A. Larsen: *Variations of box plots*. The American Statistician, 32(1):12–16, 1978, ISSN 00031305. http://www.jstor.org/stable/2683468. 33

[15] Lopes, Rogério G, Rommel N Carvalho, Marcelo Ladeira e Ricardo S Carvalho: *Predicting recovery of credit operations on a brazilian bank*. Em *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, páginas 780–784. IEEE, 2016. 79

[16] Lopes, Rogerio Gomes, Marcelo Ladeira e Rommel Novaes Carvalho: *Use of machine learning techniques in the prediction of credit recovery*. Advances in Science, Technology and Engineering Systems Journal, 2(3):1432–1442, 2017. http://astesj.com/v02/i03/p179/. 79

Apêndice A

Artigo Publicado no IEEE - *Internacional Conference on Machine Learning and Applications* (ICMLA) - 2016

2016 15th IEEE International Conference on Machine Learning and Applications

Predicting Recovery of Credit Operations on a Brazilian Bank.

Rogério G. Lopes*, Rommel N. Carvalho*†, Marcelo Ladeira* and Ricardo S. Carvalho†

*Department of Computer Science (CIC)
University of Brasilia (UnB), Brasilia, DF, Brazil
Email: rglopes@gmail.com, {mladeira,rommelnc}@unb.br

†Department of Research and Strategic Information (DIE)
Ministry of Transparency, Monitoring and Control (MTFC), Brasilia, DF, Brazil
Email: {rommel.carvalho,ricardo.carvalho}@cgu.gov.br

***Abstract*—This article presents a study conducted in a Brazilian bank, in order to assist the institution account managers in the approach to customers with loans in arrears. This approach is carried out to propose alternatives to customers return to timely payments situation, but the efficiency of this approach is small, accounting for only about 6.8% of customers. A predictive model, using classification was used to help identify customers with the most potential to return to a normal situation, reaching a 85.5% accuracy rate with the winning algorithm, Gradient Boosting Method. It was implemented in the integrated Platform H2O with R language, exploring the grid mode and parallel processing models advantages.**

I. Introduction

Since January 2015[1], we have seen a drop in credit supply and rising defaults in Brazil, resulting from decreases in economic activities and investor confidence. According to Brazil's Central Bank (BCB), credit operations for individuals are those that have shown the highest growth of default, from 5.1% in December 2014 to 6.7% in April 2016, a proportional increase of 31%. Figure 1 illustrates the gradual growth occurred in this period.

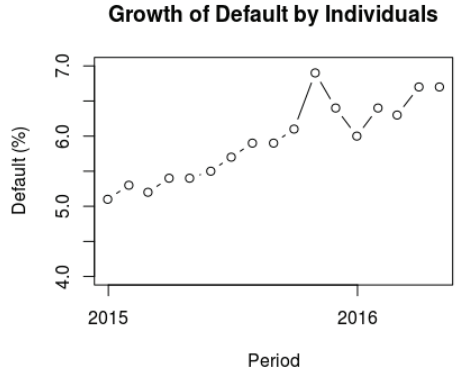

Fig. 1. Growth of Default by Individuals. Adapted from [5].

Related to this issue, there are several studies that qualify customers due to the credit risk they represent, separating them into groups of good or bad payers. However, once the bad debt occurs, there has been little research on classifying the possibility of these bad payers becoming good payers again [5].

This study was conducted on a bank with one of the highest loan portfolio of the Brazilian National Financial System (SFN). Although having a lower default rate than the numbers SFN described above, this bank also showed an increase in the event of default on individual credit operations.

Credit analysis are used for the granting of credit, using models that minimize the occurrence of default in operations. But once we found the default, the traditional action is to deny new lines of credit to these borrowers.

Due to the increase in observed default, there has been intensified actions to offer to this customers alternatives to return to normal condition of operation. The contacts were made by their account managers.

In the first months of using this approach, a major operational difficulty was detected: the small amount of customers that an account manager could contact and offer his service. In addition, it was observed that a significant part of customers did not have conditions to renegotiate his/hers operations, thus making inefficient performance of the account manager.

It is also important to note that several customers could have returned to the non-default condition, but they did not know the alternatives that were available to them.

So, this study was designed to support the account managers of this bank, in order to inform them a list of customers that are most likely to recover from default.

Therefore, the main objective of this study was to apply data mining techniques to predict which credit operations could return to a non-default situation. Models were developed using Generalized Linear Models (GLM), Gradient Boosted Methods (GBM), and Distributed Random Forest (DRF) and then compared. This comparison was made using the AUC and the PCC indicators, which will be explained on section III.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.139

The models were developed using the R language and H2O platform of data mining, considering its parallel processing capabilities. Further details on section III. [1]

This paper is organized as follows: Section II presents the credit scoring state of the art. Section III presents the methodology used in this study. Section IV presents the results given, the models learned as well their evaluating. Section V presents the conclusion and future work.

II. State of the Art

The default numbers observed in Brazil, from December 2014 to April 2016, indicate that financial institutions need a tool to support their credit granting decisions. Credit scoring models are an estimate based on the likelihood that a customer will present some undesirable behavior in the future.

Lessmann et al. [6] in a paper published in 2015, conducted a study evaluating 41 publications on the award of credit since 2006, all of them using classifiers to categorize customers as good or bad payers. These works were organized into three categories of classifiers: Individuals; homogeneous ensemble; and heterogeneous ensemble classifiers.

Six databases were used to verify the performance of each of the 41 models proposed, evaluating them from the standpoint of 6 indicators: area under the receiver operating curve (AUC), percentage correctly classified (PCC), partial Gini index, H-measure, Brier Score (BS) and Kolmogorov-Smirnov (KS).

AUC represents how well classified your data were, regardless to its distribution or misclassification costs.[8]. The PCC is an overall accuracy measure, it indicates the percentage of outcomes that were correctly classified [4][2]

At the end of this comparison, it was shown a ranking indicating that the ensemble heterogeneous algorithms had presented better overall performance. However, the difference in performance between the three categories proved to be very small. Considering the simplicity of the others algorithms, these could be used and they would provide similar results than those provided by more complex algorithms.

The detailed table I presents the results of the benchmark, bringing new performance references and recognition of new algorithms. As seen in this table, the HCES-Bag algorithm obtained the highest AUC result, while the AVG W and GASEN algorithms achieved 80.7% of PCC.

TABLE I
State of Art - Models Comparison

	Algorithm	AUC	PCC
Heterogeneous Ensemble	HCES-Bag	0.932	80.2
	AVG W	0.931	80.7
	GASEN	0.931	80.7
Homogeneous Ensemble	RF	0.931	78.9
	BagNN	0.927	80.2
	Boost	0.93	77.2
Individual	LR	0.931	70.84
	LDA	0.929	78.4
	SVM-Rbf	0.925	79.9

[1]H2O is an open source machine learning platform, available at www.h2o.ai
[2]For more details on the others indicators,[6]

III. Methodology

This study followed, where applicable, the phases of the data mining process proposed by the Cross Industry Standard Process for Data Mining (CRISP-DM): Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation.[3]

The Business Understanding phase was already covered in the introduction section I. The Deployment phase, the last one of the CRISP-DM, was not performed because this study is still in its initial state and, as later demonstrated in this paper, will still be refined before being applied to the institution's credit recovery process. The following subsections will present the others phases of the CRISP-DM.

A. Data Understanding

For this study, three data bases were used. The database (i) contained a sample of 22,764 transactions that were late in February 2016, containing variables related to the data of the contracting of credit operations, such as date of hire, time of the operation mode of the credit operation , operation risk, contracted amount, outstanding balance and amount of days in arrears, totaling 38 variables. The second database (ii) contained the status of every existing transaction in the database (i) and also the number of days with overdue operation verified in March 2016. The database (iii) was composed of 158 variables with demographic and financial information of all the clients listed in the database (i), also obtained in February 2016.

All three databases were extracted from the Data Warehouse (DW) of the institution, containing integrated and validated data without missing values and with data integrity assurance.

The variables were not individually identified in this study because the bank considered them confidential. Hence, only their categories were mentioned.

B. Data Preparation

As the data sources are from the DW, data preparation activity was reduced to developing a single database containing the 196 variables, resulting from the joint database (i), (ii) and (iii), and by creating new variables. These new variables were transformed with the following characteristics:

- Composed primary keys: have been identified and each of these keys has been transformed into a single variable of type factor.
- Date type fields: the ones in which the interest was in the period of time between the date the event occurred and the time being of this study (February 2016) were transformed into numerical variables, representing that period time.
- Target variable: a binomial categorical variable was created, containing the value 1 for the operations that reduced the delay and 2 for those that have maintained or increased the number of days overdue, comparing databases (ii) and (i).

C. Modeling

Three models were developed to be compared and then one was chosen as the best model to be used in generating support information to account managers. At first, the models were being processed in the software R. However, considering the large amount of variables (196), processing the data was taking too long and that was compromising the efficiency of the study. So, the platform H2O was used, integrated with R, to explore the grid mode and parallel processing models. The grid allowed to combine different parameters to build different models. The parallel processing allowed those models to be built at the same time. This efficiency increase made it possible to build more models, with better tuning parameters.

By using the R language, integrated with H2O platform, the following algorithms were used:

- GLM - Generalized linear Modeling
- GBM - Gradient Boosting Method
- DRF - Distributed Random Forest

Generalized Linear Models are similar to linear regression, only it's more flexible for it doesn't requires normal distribution to the errors. It estimates models for outcomes from the Exponential family and it is used for both regression and classification. fits really well to large data sets. GLM fits really well to large data sets and is very popular because of its easy interpretation and its speed, even when used for data sets with great number of columns. [7]

The Gradient Boosting Machine is used for predictive results for regression or classification. It is an ensemble of tree models and provides considerably accurate results. GBM applies weak classification algorithms to incrementally changing data, creating that way a series of decision trees. It is robust, not implying any distribution to the data, and because of that it is considered one of the best choices for many users for it requires little adjustments. [7]

Distributed Random Forest, just as GBM, is an ensemble of tree models, which each tree is de-correlated from all other trees.[7]

D. Evaluation

The evaluation of the models was performed using the two indicators analysis: Area Under Roc (AUC) and Percentage True Correctly Classified (PTCC). PTCC is a adaptation of PCC indicator (II), considering only the percentage of true positives classified.

IV. Results

In this section, we will present the results in each phase carried out during the study, as mentioned in Section III.

A. Data Preparation

Using the feature engineering technique has identified the need to create three new variables were performed and the following steps to generate the database that was used for modeling:

- Database (i) - A new variable was created to replace two variables representing the mode of the original contract for the operation, which was represented by a composed primary key in the operational system of the institution. Furthermore, the variable representing the date of the contract was transformed into a variable to indicate the lifetime of the contract.
- Database (iii) - A new variable was created to replace two variables that together represented the profession pursued by the individual customer contracting the operation.
- Target Variable - This variable was created to indicate that the operation had returned to normal after 30 days. This variable was called Delay Reduction Index (DRI), containing the value 1 to indicate that there was a reduction in delay and 0, to indicate that the delay had increased or remained the same.
- Final Database - Carried the junction of databases (i), (iii) and the target variable, with 22,764 records and 199 variables.
- The database was partitioned into 2 parts, one for training and the other for validation at a ratio of 80:20. Table II shows the composition of each partition.
- The initial analysis of the data, it showed that out of the total of 22,764 operations that had late payments, only 1,548 returned to a regular situation, which represents only 6.8% of the operations. That required a special attention in development of the predictive model, considering it features a case of imbalanced classes. Brown [2] has demonstrated that GBM and DRF perform well even in those cases. On the other hand, Verbeke [8] affirms that classification techniques perform best with balanced classes, over/under-sampling the data. It was decided to under-sample the training partition of the data, not altering the validation partition.

TABLE II
PARTITIONING THE DATABASE

Partition	DRI=1	DRI=2	Rows
Train	1.232	17.013	18.245
Test	316	4.203	4.519
Total	1.548	21.216	22.764

B. Modeling

Three predictive models were developed using the H2O platform accessed through the R language, using the algorithms GLM, GBM, and DRF. The results are detailed in the following sections.

1) GLM - Generalized Linear Modeling: This algorithm was implemented with 10-folds cross validation. The use of this validation technique was reproduced in other algorithms used.

The GLM algorithm obtained AUC = 0.956881 with 10-fold cross-validation. The PTCC obtained by analyzing the successes of the class of interest (DRI = 1) reached 63.63%, as shown in Figure 2 and Table III.

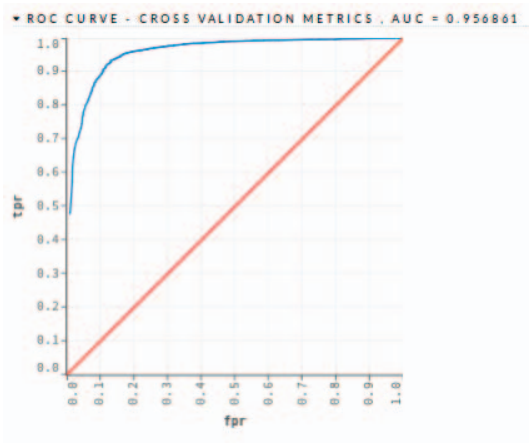

Fig. 2. GLM - AUC

TABLE III
GLM - CONFUSION MATRIX

	1	2	Error	Rate	PTCC
1	784	448	0.363636	448/1232	63.63%
2	284	16729	0.016693	284/17013	
Totals	1068	17177	0.040121	732/18245	

C. GBM - Gradient Boosting Method

The first results showed that GBM algorithms had a much performance than the others. Then, a grid of parameters was chosen and applied to it. The parameters used were:

- maximum trees: 100, 500 and 1.000.
- maximum depth: 5, 7 and 10.
- stopping tolerance: 0.001

The best model was built with 10-folds cross validation, maximum 500 trees and 7 maximum depth of them.

The GBM algorithm obtained AUC = 0.982650 with 10-fold cross-validation and PTCC for the class of interest reached 84.65%, as shown in Figure 3 and Table IV.

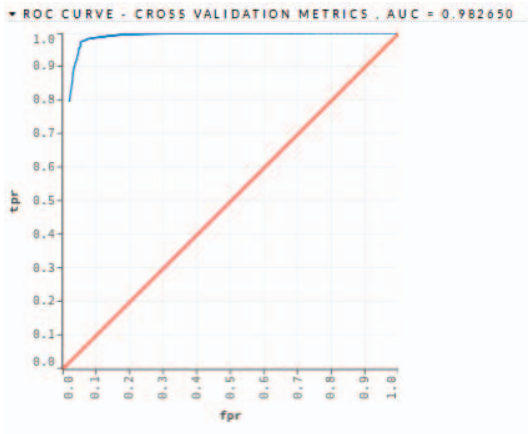

Fig. 3. GBM - AUC

TABLE IV
GBM - CONFUSION MATRIX

	1	2	Error	Rate	PTCC
1	1043	189	0.153409	189/1232	84.65%
2	101	16912	0.005937	101/17013	
Totals	1144	17101	0.015895	290/18245	

D. DRF - Distributed Random Forest

This algorithm was implemented with 10-folds cross validation, maximum 500 trees and 7 maximum depth of them.

The GBM algorithm obtained AUC = 0.978096 with 10-fold cross-validation and PTCC for the class of interest reached 61.60%, as shown in Figure 4 and Table V.

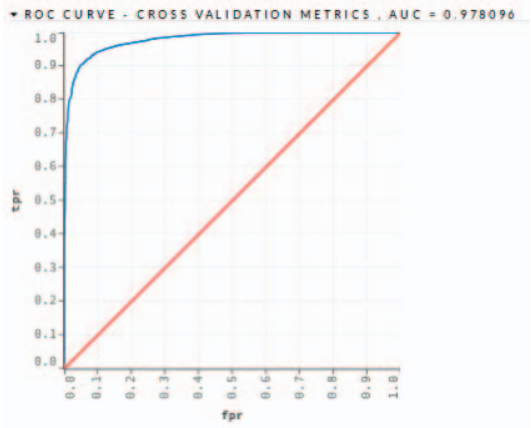

Fig. 4. DRF - AUC

TABLE V
DRF - CONFUSION MATRIX

	1	2	Error	Rate	PTCC
1	759	473	0.383929	473/1232	61.60%
2	123	16890	0.007230	123/17013	
Totals	882	17363	0.032666	596/18245	

E. Comparison of Results

All three models generated had a high evaluation result of the ROC curve. However, considering the classes were imbalanced, we cannot use this indicator alone. The large number of hits from the majority class, which is not the class of interest, leads to biased results.

Therefore, it is necessary to evaluate the second indicator, the PTCC, pointing accuracy level in the class of interest. In assessing this aspect, it is possible to realize a greater variation between the three models, ranging from 61% to 84%, with the best result obtained by the GBM algorithm. Table VI shows the comparison of the results obtained by the three models, where it is clear that the GBM algorithm achieved better performance in all metrics used.

TABLE VI
COMPARING MODELS

Algorithm	AUC	PTCC (%)
GLM	0.956881	63.63
GBM	**0.983650**	**84.65**
DRF	0.978096	61.60

V. CONCLUSION

The aim of this study was to identify those delinquent clients that possessed the highest probability of short-term recovery, to support the activities of Account Managers, increasing the efficiency of their approach with customers.

Although it was inspired by the study of the state of the art of credit score models, if failures occur in the classification, there is no financial penalty, since the credit operations are already late. In theory, the errors only decrease the performance of Account Managers.

In comparison between the algorithms, the GBM showed a better performance in both indicators calculated, making the candidate to be chosen for implementation and integration with bank's operating systems.

The percentage of customers who can return to timely payments proved to be close to 6.8%. By using the proposed predictive model, the account managers may increase the efficiency of the approaches made to customers, considering the PTCC index of 85.5%, bringing gains to the credit recovery activity.

The performance indicators obtained in this study cannot be directly compared with Lessman [6], since different databases were used in each study. However, because they have close performances in AUC and PCC indicators, we can say that the model is suitable for use in the bank.

A. Future Works

This was just an initial study, which can be further enhanced with the use of other predictive modeling techniques, such as using ensemble learning, both homogeneous and heterogeneous.

In addition, there is a chance that the behavior of lower default rates have a seasonal behavior, over a year, due to situations that occur with expenses that are held on fixed dates a year, such as taxes, school fees and also seasonal variations in income receipts. So, it is indicated the continuity of this work developing predictive models for different times over a year.

REFERENCES

[1] Poltica Monetria e Operaes de Crdito do SFN.

[2] Iain Brown and Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446 – 3453, 2012.

[3] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[4] Karel Dejaeger, Frank Goethals, Antonio Giangreco, Lapo Mola, and Bart Baesens. Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, 218(2):548 – 562, 2012.

[5] Sung Ho Ha. Behavioral assessment of recoverable credit of retailer's customers. *Inf. Sci.*, 180(19):3703–3717, October 2010.

[6] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[7] The H2O.ai team. *h2o: R Interface for H2O*, 2015. R package version 3.1.0.99999.

[8] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211 – 229, 2012.

Apêndice B

Artigo Estendido Publicado no *Advances in Science, Technology and Engineering Systems Journal* (ASTESJ) - 2017

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1432-1442 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Use of machine learning techniques in the prediction of credit recovery

Rogerio Gomes Lopes[*,1], Marcelo Ladeira[2], Rommel Novaes Carvalho[2]

[1]*Bank of Brazil, IT Department, Brazil, rglopes@bb.com.br*

[2]*University of Brasilia, Department of Computer Science, Brasilia, Brazil, mladeira@unb.br, rommel.carvalho@gmail.com*

ARTICLEINFO

Article history:
Received: 04 June, 2017
Accepted: 28 July, 2017
Online: 10 August, 2017

Keywords :
machine learning
data mining
credit recovery
h2o.ai

ABSTRACT

This paper is an extended version of the paper originally presented at the International Conference on Machine Learning and Applications (ICMLA 2016), which proposes the construction of classifiers, based on the application of machine learning techniques, to identify defaulting clients with credit recovery potential. The study was carried out in 3 segments of a Bank's operations and achieved excellent results. Generalized linear modeling algorithms (GLM), distributed random forest algorithms (DRF), deep learning (DL) and gradient expansion algorithms (GBM) implemented on the H2O.ai platform were used.

1 Introduction

This paper is an extension of the work originally presented at the International Conference on Machine Learning and Application (ICMLA 2016) [1], which presented the first results of a Brazilian bank research to reduce its losses with defaulting clients. That study covered only a sample of 22.764 transactions, representing a homogeneous group of bank customers. We extend our previous work by adding all operations from individual costumers which were in arrears in July 2016.

The Figure 1 shows that there was a slight decrease in the number of debtors in June 2016, but increased again in the following months.

The Bank had nearly 54 million active credit agreements with individuals at the end of July 2016. Of this amount, approximately 8.6 million were delayed for 15 days or more, accounting for 15.9% of the contracts. These delinquent contracts amounted to more than R$20.8 billion (US$6.4 billion in July 2016), accounting for approximately 5.8% of the Bank's individuals loan portfolio, an increase of 1.2 percentage points over December 2014. That is, in 21 months the financial volume of overdue loans contracted by individuals increased by 26%.

The Brazilian Central Bank (BACEN) regulation requires financial institutions to classify their credit operations and perform a Provision for Doubtful Accounts (PDA), according to a risk classification. The main criteria for the classification is the number of days in arrears of each individual credit agreement. The Table 1 shows the days-in-delay ranges considered to determine a risk classification and therefore the minimum percentage PDA that financial institutions must reserve. As an operation increases the number of days in arrears, there is a non-linear increase of PDA, which may allocate 100% of the outstanding balance of the contract. For example, an operation with a debit balance of R$ 1,000, with 15 days in arrears, must reserve a minimum provision of R$ 10. The amount of the provision may reach R$ 1,000 if the arrear reach 180 days.

Table 1: Days in arrears x Provision

Days in arrears	Minimum Risk	PDA %
15-30	B	1
31-60	C	3
61-90	D	10
91-120	E	30
121-150	F	50
151-180	G	70
over 180	H	100

At the time of the credit granting, financial institutions assume the credit risk and make the corresponding provisions in accordance with the current Central Bank regulation. Acting in this way, in a possible default of the customer, the financial institution and the stability of the financial system will be protected. However, as a customer delays its operations, the natural reaction of financial institutions is to restrict credit to them, increasing the chances of these

[*]Rogerio Gomes Lopes, Braslia, Brazil, rglopes@bb.com.br.

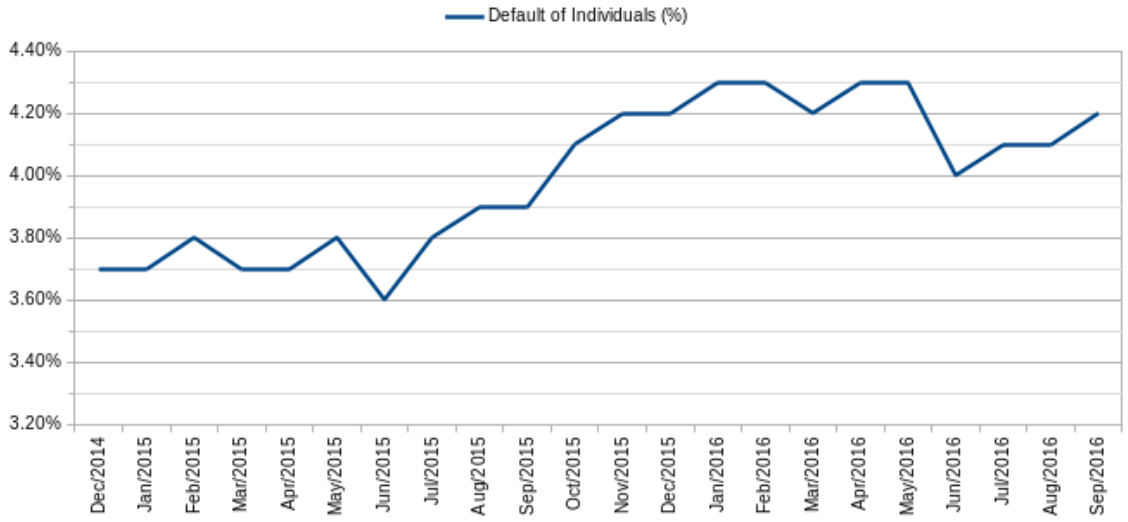

Figure 1: Default of individuals.

customer's evasion to other institutions, since they will not be able to carry out new credit operations with the original institution.

With the increase in delinquency, a mobilization of account managers of the bank began in order to mitigate the evasion of its clients by approaching the customer in arrear and proposing alternatives that could fix the delayed payments. Hence, solving the default situation and the possible loss of the customer of its portfolio, as well as reducing the financial amount allocated to (PDA).

Provided that the selection of the clients is a time and resource consuming task, the main objective of this study was to apply machine learning techniques to predict the recovery probability of credit transactions, providing a list of delinquent clients with the greatest potential for regularization of their operations.

Models were developed using Generalized Linear Models (GLM), Gradient Boosted Methods (GBM), Distributed Random Forest (DRF) and Deep Learning (DL)[1] . The models were compared using the recall indicator, which will be explained on section 3. The models were developed using the R language and H2O machine learning platform, considering its parallel processing capabilities. Further details on section 3. [2]

This paper is organized as follows: Section 2 presents the credit scoring state of the art. Section 3 presents the methodology used in this study. Section 4 presents the modeling and evaluation of the generated models for each method. Section 5 presents the conclusion and future works.

2 State of the Art

The default numbers observed in Brazil, from December 2014 to September 2016, indicate that financial institutions need a tool to support their credit granting decisions. Although there are several studies to identify the customer credit risk, qualifying them as good or bad payers, helping to make a decision to grant credit, there is few research studying the credit recovery, when the delinquency occurs. [2]

In [3], the author conducted a study evaluating 41 publications on the award of credit since 2006, all of them using classifiers to categorize customers as good or bad payers. Those works were organized into three categories of classifiers: individuals; homogeneous ensemble; and heterogeneous ensemble classifiers. Most of the algorithms used were implemented through logistic regression and decision trees, with their use of boosting, bagging and forest variants.

The Table 2 lists the eight datasets that were used in [3] to verify the performance of each of the 41 models proposed, evaluating them from the standpoint of 6 indicators: Area Under the Receiver Operating Curve (AUC), percentage correctly classified (PCC), partial Gini index, H-measure, Brier Score (BS) and Kolmogorov-Smirnov (KS).

Table 2: Datasets used in [3].

Name	Samples	Features	Debtors %
AC	690	14	44.5
GC	1000	20	30.0
Th02	1225	17	26.4
Bene 1	3123	27	66.7
Bene 2	7190	28	30.0
UK	30000	14	4.0
PAK	50000	37	26.1
GMC	150000	12	6.7

[1] Documentation available at http://docs.h2o.ai/h2o/latest-stable/index.html
[2] H2O is an open source machine learning platform, available at www.h2o.ai

In [4], the author presents AUC as an indicator that represent how well classified were the data, independent of its distribution or misclassification costs. PCC is an overall accuracy measure that indicates the percentage of outcomes that were correctly classified.[5]

A score was assigned to each algorithm, referring to the classification received in the comparison between them within the same performance measure . For example, the algorithm K-means was in 12th place considering the AUC indicator, while the KNN was in 29th place. Thus, the scores attributed to them were 12 and 29, respectively. Then, the algorithms were ordered by the average of all metrics, where the 1st place were the algorithm that obtained the lowest score.

The heterogeneous multi-classifiers presented a better performance, although the performance between the three categories was very similar.

The Table 3 presents the results of the benchmark, indicating that the HCES-Bag algorithm obtained the highest AUC result, while the AVG-W and Gasen algorithms reached 80.7% of the PCC.

Table 3: State of Art - Models Comparison - Adapted from [3]

	Algorithm	AUC	PCC
Heterogeneous Ensemble	HCES-Bag	0.932	80.2
	AVG W	0.931	80.7
	GASEN	0.931	80.7
Homogeneous Ensemble	RF	0.931	78.9
	BagNN	0.927	80.2
	Boost	0.93	77.2
Individual	LR	0.931	70.84
	LDA	0.929	78.4
	SVM-Rbf	0.925	79.9

3 Methodology and Infrastructure Setup

This section presents the methodology used in this study, which was segmented in stages according to the phases proposed by CRISP-DM [6]. The result of each phase is described in the next Section.

Training model environment - The models were trained on the H2O.ai platform, in a cluster formed by 5 virtual machines on the same subnet and with the same configuration. Their operating system was Red Hat Enterprise Linux 6.8 64 bits, with 34 cores and 80 GB of RAM. It were used H2O.ai version 3.10.4.5 and R version 3.3.0. It were allocated 44 GB of RAM and all cores of each machine, reaching a total of 170 cores and 220 GB of RAM.

The training dataset consisted of about 40 million copies, requiring a robust platform to be made available for the processing of this data.

The Figure 2 shows the CPU meter of the H2O.ai cluster in action at the moment of the training models. It shows the percentage of use of the processors of each machine, identified by the final number of its IP address (174 to 178) and the port number where the service was running (54321). The intensive use of the 170 available cores shown in the Figure 2 reinforces the need for a robust platform.

Each vertical bar represents 1 core and the colors represent the type of process executed: idle time (blue), user time (green) and system time (red).

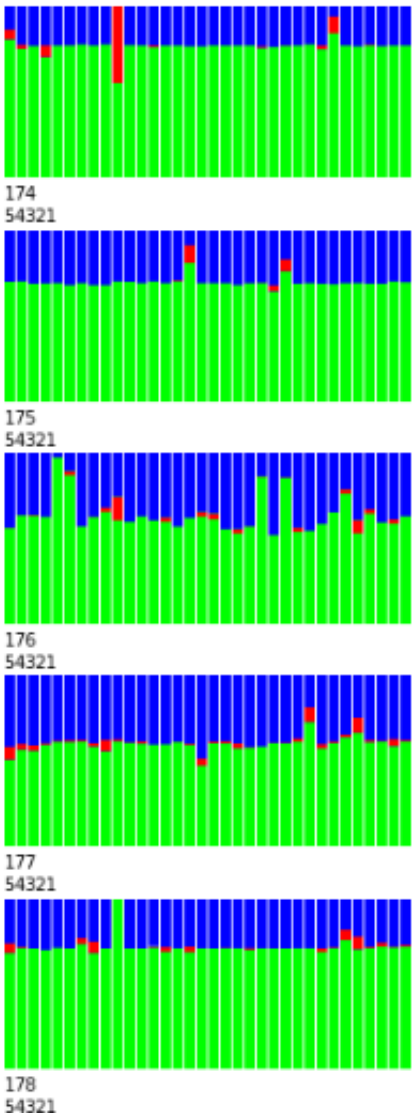

Figure 2: Cluster H2O in action

4 Results

In this sections, the results of the CRISP-DM phases are detailed: Data Understanding, Data preparation, Modeling, Evaluation and Implementation.

4.1 Data Understanding

The dataset was obtained by the extraction of information from legacy systems and customers relationship data marts. It has information about customers accounting,demographic and financial data. The dataset had 28 features and 1 label that indicates the recovery of the respective credit operation. The

Tables 4 and 5 present these 28 characteristics organized by categorical and numerical features.

Table 4: Numeric features

Features	Description
V1	Number of days of delinquency.
V2	Number of days remaining for the end of the contract.
V3	Contract value.
V4	Amount of the outstanding balance.
V5	Amount PDA provisioned for the contract.
V6	Percentage loss expected for the contract.
V7	Quantity of products owned by the customer.
V8	Time of customer relationship with the Bank.
V9	Customer age.
V10	Customer income.
V11	Customer total contribution margin amount.
V12	Value of Gross Domestic Product per capita

Table 5: Categorical features

Features	Description
VC1	Customer portfolio type.
VC2	Customer behavioral segment.
VC3	Product.
VC4	Product modality.
VC5	Structured operation indicator.
VC6	Management level that approved the operation.
VC7	Transaction risk credit.
VC8	Range of past delays.
VC9	PDA lock indicator.
VC10	Customer relationship with the bank.
VC11	Client instruction level.
VC12	Customer gender.
VC13	Nature of customer occupation.
VC14	Customer registration status.
VC15	Customer's age group.
VC16	Age group of relationship time.

For the data understanding, the analysis began in July 2016 containing all credit operations contracts, regardless of the contracted product, with more than 14 days in arrears. In addition, transactions with the highest risk were considered as already lost contracts by our business specialists and removed from our dataset.

For definition of the label, the delay reduction indicator, the following operation was performed, considering that the data of the delayed operations were used in July 2016:

- Delay Reduction Indicator = 1, for all transactions that showed a reduction in the number of days overdue in the subsequent month, that is, in August 2016, or that their debit balances have been reduced.

- Delay Reduction Indicator = 0, otherwise, that is, presenting a delay or debit balance in August 2016 equivalent to or greater than that observed in July 2016.

The Table 6 presents the summary of transactions in the month of July 2016, which resulted in a base with 4,514,029 contracts. Of this total, only 271,193 (6.01%) were recovered.

Table 6: Dataset July 2016

Not recovered	Recovered
4,242,836	271,193
93.99%	6.01%
Samples	4,514,029

The bank has several strategies for credit recovery, according to the customer profile and the category of the credit operation, grouping them with distinct trading rules. Existing segments are divided into massive and individual strategies. Massive strategies are implemented for segments that have a known behavior pattern, whereas individual strategies cover operations that have atypical or special characteristics which require a case-by-case analysis to perform a collection and recovery.

Based on this information, the dataset was splitted into segments compatible with the institution's recovery strategies, grouping similar products and customer segments with characteristics in common removing from the study the segments that have an individualized trading strategy. The Table 7 lists the 11 segments that will be worked on in this study, in addition to the Individualized Strategy segment, which was removed from the study.

4.2 Data Preparation

In this study, the analysis were performed only in the first 3 segments, Mortgage Loan I, II and III. The remaining segments are in the final analysis phase and will be presented at a later time.

Then, the data preparation was started, analyzing each one of the segments, preparing the data sets for the modeling phase.

The Tables 8, 10 and 12 present the summary of descriptive analysis of the numerical features of segments Mortgage Loan I, II and III, respectively. In these tables the data of quartiles and Kendall's Tau [7] of each feature are presented.

The Tables 9, 11 and 13 present the summary of the descriptive analysis of the categorical variables, listing the Kendall's Tau and the number of levels of each feature.

Table 7: Credit Operations Segments

Segment	Credit Recovered				Samples
	No		Yes		
	Qty	%	Qty	%	
Mortgage Loan I	41,398	70.45	17,365	29.55	58,763
Mortgage Loan II	400	73.94	141	26.06	541
Mortgage Loan III	3,537	78.11	991	21.89	4,528
Vehicle Financing I	12,115	87.90	1,667	12.10	13,782
Vehicle Financing II	32,357	86.63	4,993	13.37	37,350
Agribusiness	258,618	98.84	3,021	1.16	261,639
Social Business	137,474	93.53	9,504	6.47	146,978
Credit Card I	17,124	98.92	187	1.08	17,311
Credit Card II	454,864	98.56	6,661	1.44	461,525
Other Operations Income I	186,572	96.53	6,714	3.47	193,286
Other Operations Income II	2,668,890	92.96	201,977	7.04	2,870,867
Individualized Strategy	429,487	95.98	17,972	4.02	447,459

Table 8: Mortgage Loan I - Numerical Features

Feature	Min	1QT	Median	Avg	3QT	Max	Kendall's Tau
V1	15	20	51	87.43	112	624	-0.29
V2	0	10,180	10,420	10,290	10,670	11,620	-0.03
V3	0	0.1	0.13	0.17	0.27	0.67	-0.02
V4	0	1	2	3.94	5	26	0.06
V5	17	25	29	31.28	36	73	0.03
V6	1	3	4	4.45	5	37	0.08
V7	14,790	74,400	87,460	86,270	97,470	164,800	0.01
V8	-124,400	-390.8	156.7	-1,397	256.3	183,400	0.38
V9	0	1,349	3,221	3,712	5,525	46,040	0.04
V10	0	888.1	2,670	19,090	20,960	173,700	-0.17
V11	0	1,586	1,700	1,877	2,000	20,000	0.03
V12	0.68	74,920	88,720	87,250	99,510	173,500	0.00

4.3 Modeling

For each dataset, 4 predictive models were elaborated, using the H2O platform integrated to the R, using the algorithms Generalized Linear Models (GLM), Gradient Boosting Method (GBM), Random Forest (DRF) and Deep Learning (DL). The first three algorithms were chosen because they represent the techniques most used in the calculation of credit risk, which performs a classification task very similar in [8]. The algorithm DL was used to verify its behavior in a knowledge area not yet explored, but with expectation of good suitability due to the use of a great amount of variables. [9]

The datasets of the Mortgage Loan I and III segments were splitted into 3 parts: 70% for training, 20% for validation and 10% for testing. Due to the small number of observations in the Mortgage Loan II, this dataset was splitted only in training and validation in a proportion of 80% and 20%, respectively. The next subsections present the evaluation results for each segment.

4.3.1 Mortgage Loan I

- GLM - This algorithm obtained an AUC = 0.7774755 and a PCC of 66.53%, as shown in the Figure 3 and in the Table 14

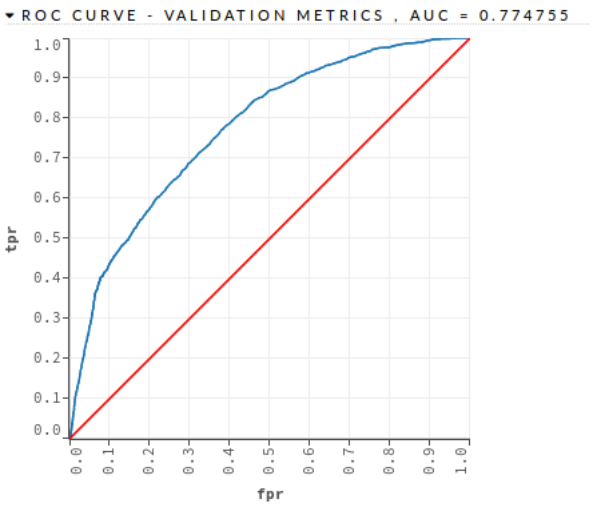

Figure 3: Mortgage Loan I - GLM - Validation Dataset AUC

Table 9: Mortgage Loan I - Categorical Features

Feature	Kendall's Tau	Number of levels
VC1	0.23	6
VC2	0.02	5
VC3	-0.09	3
VC4	-0.21	9
VC5	0.03	12
VC6	0.06	7
VC7	-0.01	2
VC8	0.03	5
VC9	-0.04	16
VC10	0.00	4
VC11	0.05	8
VC13	-0.21	9
VC14	0.01	4
VC15	-0.02	18
VC16	0.00	2

Table 10: Mortgage Loan II - Numerical Features

Feature	Min	1QT	Median	Avg	3QT	Max	Kendall's Tau
V1	15	21	48	89	113	507	-0.15
V2	0	1,626	2,928	3,037	4,076	6,776	-0.14
V3	0	0	0	0	0	1	0.09
V4	2	6	6	10	13	31	0.03
V5	30	42	48	49	55	85	0.02
V6	1	4	6	7	9	36	-0.04
V7	4,400	28,000	45,000	59,560	70,560	240,000	-0.05
V8	-31,770	-148	91	-650	371	25,070	0.30
V9	0	4,990	6,190	6,663	9,099	15,260	0.08
V10	0	173	650	8,744	10,230	116,000	-0.20
V11	0	1,598	2,965	5,082	5,553	128,900	-0.11
V12	0	9,316	20,500	36,670	47,120	215,200	-0.14

- DRF - This algorithm was implemented with 500 trees and a maximum depth of 7. The DRF algorithm obtained an AUC = 0.880589 and a PCC = 75.85%, as shown in the Figure 4 and in the Table 14

- DL - Deep Learning This algorithm was implemented with 2 hidden layers with 200 neurons each one. The DRF algorithm obtained an AUC = 0.898203 and a PCC = 79.22%, as shown in the Figure 5 and in the Table 14.

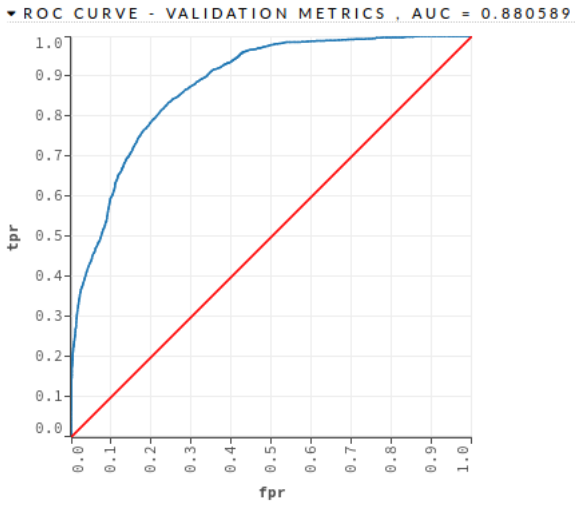

Figure 4: Mortgage Loan I - DRF - Validation Dataset AUC

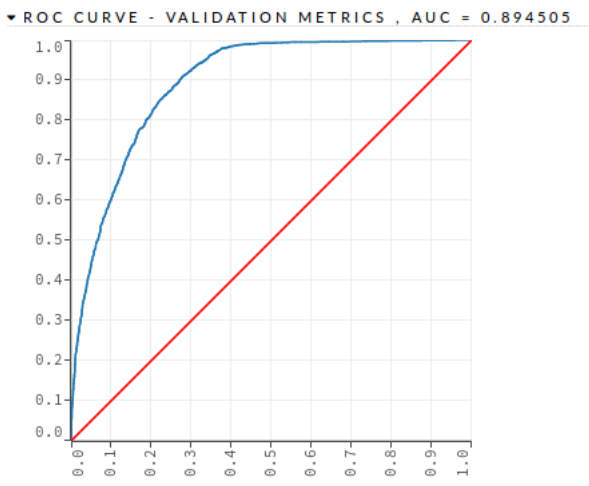

Figure 5: Mortgage Loan I - DL - Validation Dataset AUC

Table 11: Mortgage Loan II - Categorical Features

Features	Kendall's Tau	Number of levels
VC1	0.11	4
VC2	0.13	3
VC4	-0.19	8
VC5	0.01	10
VC6	0.03	6
VC7	-0.04	2
VC8	0.10	5
VC9	0.06	10
VC11	-0.15	5
VC13	-0.19	8
VC14	0.10	4
VC15	-0.06	13

Table 12: Mortgage Loan III - Numerical Features

Feature	Min	1QT	Median	Avg	3QT	Max	Kendall's Tau
V1	15	30	81	131	181	511	-0.31
V2	0	4,876	7,530	6,616	8,203	10,910	-0.01
V3	0.00	0.02	0.05	0.06	0.07	0	0.00
V4	0	5	9	11	15	54	-0.02
V5	20	35	43	44	52	78	-0.06
V6	2	6	8	10	11	71	0.05
V7	20,000	100,000	142,500	188,700	213,800	3,000,000	-0.07
V8	-257,600.00	-5,476.00	-930.00	-9,087.00	257.70	199,600	0.36
V9	0	2,769	4,660	4,968	6,485	46,040	0.01
V10	0	3,317	14,200	51,540	53,470	1,212,000	-0.20
V11	0	2,280	5,542	10,700	11,130	337,600	-0.01
V12	250	88,900	134,500	177,500	203,200	3,084,000	-0.08

- GBM - This algorithm was implemented with 500 trees and a maximum depth of 7. The GBM algorithm obtained an AUC = 0.988574 and a PCC = 93.90%, as shown in the Figure 6 and in the Table 14

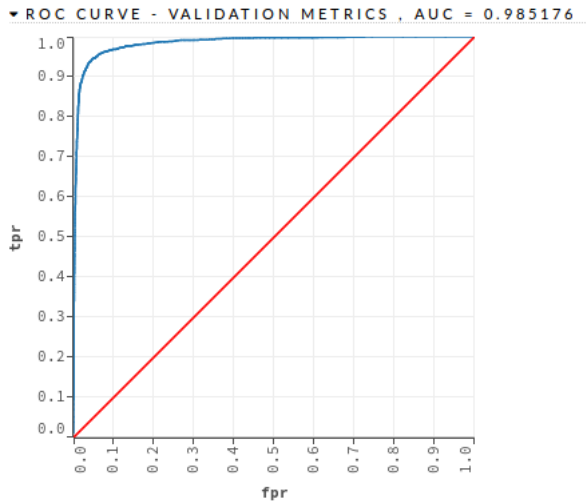

Figure 6: Mortgage Loan I - GBM - Validation Base AUC

4.3.2 Mortgage Loan II

Because of the small number of records, this dataset was splitted only in training and testing, in the ratio of 80:20, and validation was performed through cross validation with 10 folds.

- GLM - This algorithm obtained an AUC = 0.848474 and a PCC = 65.51%, as shown in the Figure 7 and in the Table 15.

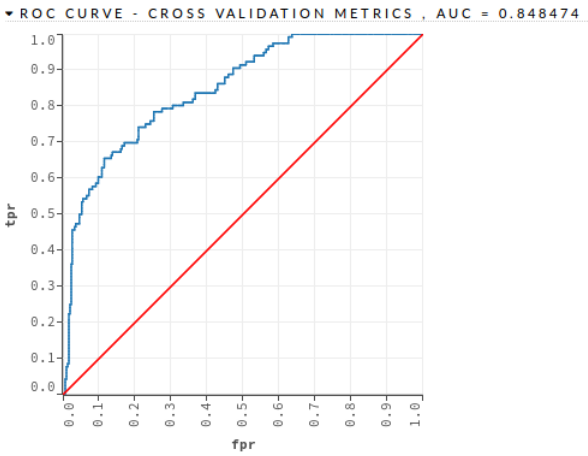

Figure 7: Mortgage Loan II - GLM - Validation Dataset AUC

Table 13: Mortgage Loan III - Categorical Features

Feature	Kendall's Tau	Number of levels
VC1	0.16	6
VC2	0.00	4
VC3	-0.24	2
VC4	-0.21	33
VC5	-0.06	12
VC6	-0.02	7
VC7	-0.03	2
VC8	0.00	5
VC9	-0.02	16
VC10	-0.10	5
VC11	0.03	7
VC12	0.00	2
VC13	-0.21	9
VC14	0.00	5
VC15	-0.04	17
VC16	0.17	2

Table 14: Mortgage Loan I - Confusion Matrix

Algorithm		0	1	Err %	PCC
GLM	0	5003	3050	37.87	
	1	776	2603	22.96	
	Total	5779	5653	33.46	66.52
DRF	0	6638	1415	17.57	
	1	816	2563	24.14	
	Total	7454	3978	19.51	75.85
DL	0	6290	1763	21.89	
	1	517	2862	15.39	
	Total	6897	4625	19.94	79.22
GBM	**0**	**7770**	**283**	**3.51**	
	1	**238**	**3141**	**7.04**	
	Total	**8008**	**3424**	**4.55**	**93.90**

Table 15: Mortgage II - Confusion Matrix

Algorithm		0	1	Err %	PCC
GLM	0	270	35	11.47	
	1	40	76	34.48	
	Total	310	111	17.81	65.51
DRF	0	302	3	0.98	
	1	17	99	14.65	
	Total	319	102	4.75	85.34
DL	**0**	**297**	**8**	**2.62**	
	1	**10**	**106**	**8.62**	
	Total	**307**	**114**	**4.27**	**91.37**
GBM	0	301	4	1.31	
	1	16	100	13.79	
	Total	317	104	4.75	86.20

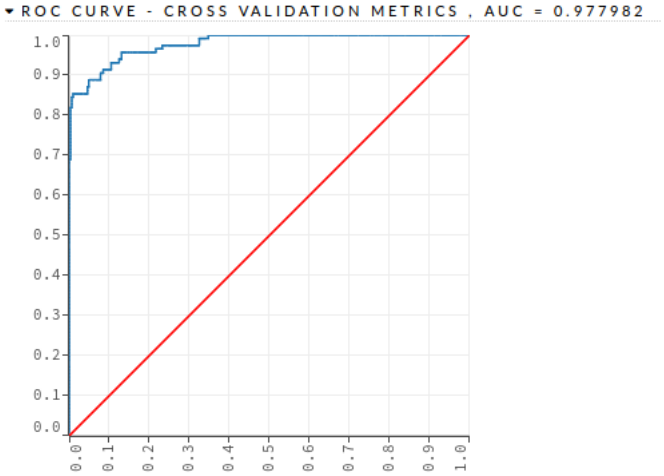

Figure 8: Mortgage Loan II - DRF - Validation Dataset AUC

- DRF - This algorithm was implemented with 500 trees and a maximum depth of 7. The DRF algorithm obtained an AUC = 0.977982 and a PCC = 93.10%, as shown in the Figure 8 and in the Table 15

- DL - This algorithm was implemented with 2 hidden layers with 200 neurons each one. The DRF algorithm obtained an AUC = 0.956868

and a PCC = 91.37%, as shown in the Figure 5 and in the Table 15.

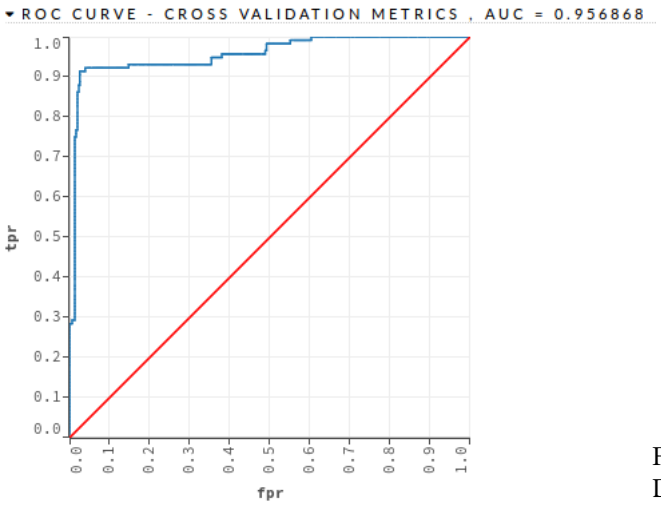

Figure 9: Mortgage Loan II - DL - Validation Dataset AUC

- GBM - This algorithm was implemented with 500 trees and a maximum depth of 7. The GBM algorithm obtained an AUC = 0.972640 and a PCC = 86.20%, as shown in the Figure 10 and in the Table 15

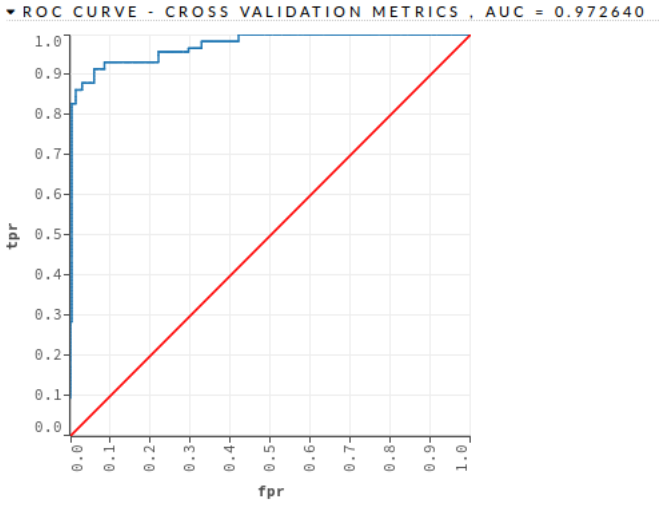

Figure 10: Mortgage Loan II - GBM - Validation Dataset AUC

4.3.3 Mortgage Loan III

- DRF - This algorithm was implemented with 500 trees and a maximum depth of 7. The DRF algorithm obtained an AUC = 0.950718 and a PCC = 83.51%, as shown in the Figure 11 and in the Table 16

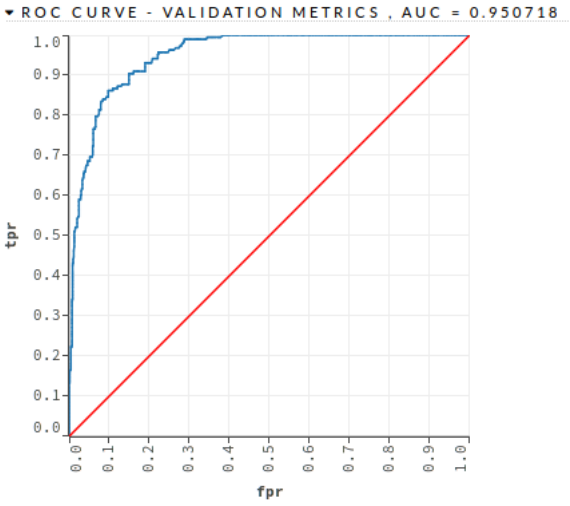

Figure 11: Mortgage Loan III - DRF - Validation Dataset AUC

- DL - This algorithm was implemented with 2 hidden layers with 200 neurons each one. The DRF algorithm obtained an AUC = 0.939082 and a PCC = 78.20%, as shown in the Figure 12 and in the Table 16.

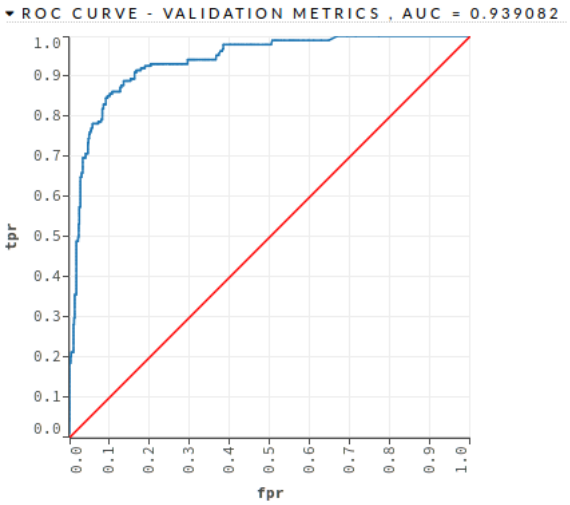

Figure 12: Mortgage Loan III - DL - Validation Dataset AUC

- GBM - This algorithm was implemented with 500 trees and a maximum depth of 7. The GBM algorithm obtained an AUC = 0.955728 and a PCC = 98.93%, as shown in the Figure 13 and in the Table 16

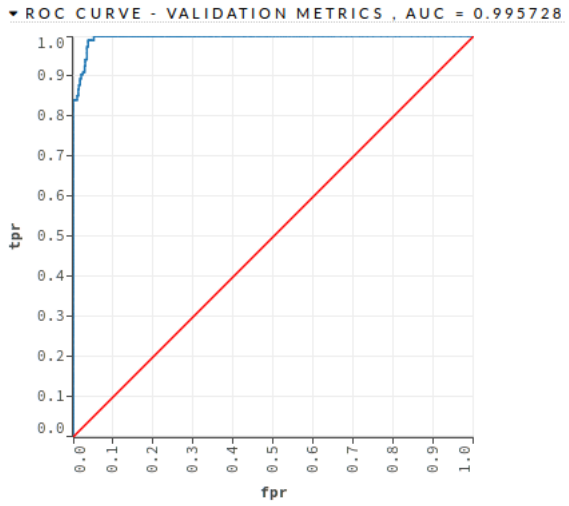

Figure 13: Mortgage Loan III - GBM - Validation Dataset AUC

- GLM - This algorithm obtained an AUC = 0.814560 and a PCC = 60.10%, as shown in the Figure 14 and in the Table 16

Table 16: Mortgage Loan III - Confusion Matrix

Algorithm	0	1	Err %	PCC
GLM	601	95	13.64	
	75	113	39.89	
	676	208	19.23	60.10
DRF	640	56	8.04	
	31	157	16.48	
	671	213	9.84	83.51
DL	656	40	5.74	
	41	147	21.80	
	697	187	9.16	78.20
GBM	**670**	**26**	**3.73**	
	2	**186**	**1.06**	
	672	**212**	**3.16**	**98.93**

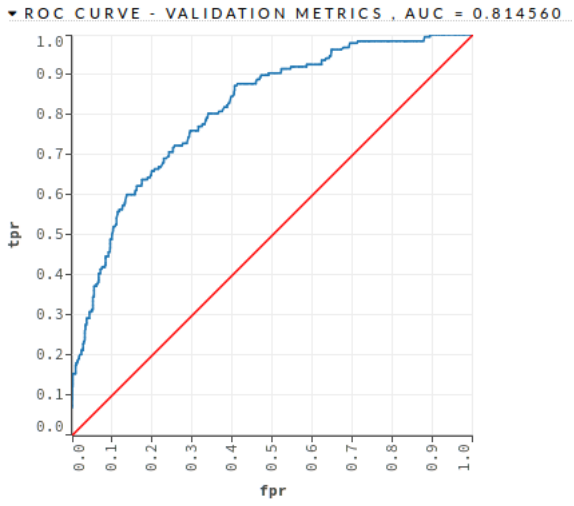

Figure 14: Mortgage Loan III - GLM - Validation Dataset AUC

5 Conclusion

The main objective of this study was to apply machine learning techniques to predict the probability of recovery of credit transactions, providing a list of defaulting clients with greater potential for regularization of their operations.

Studies were carried out on 3 segments of credit operations, which have different recovery strategies, the Mortgage Loan segments I, II and III. With the machine learning, it was possible to elaborate predictive models with great contribution to assist the Managers in the approach to their clients with operations in arrears.

Mortgage Loan I - The model with the highest recall was obtained with the GBM algorithm. In a total of 11,342 contracts in default, there were 3,424 contracts recovered. The model was able to correctly predict 3,141 contracts, reaching a recall of 92.86%. Using the prioritization list generated by the model, the work of Bank Managers would be more assertive. In addition, the model correctly predicted 7,770 (97.02%) contracts, out of 8,008 contracts that would not be recovered.

Mortgage Loan II - The model with the highest recall was obtained with the DL algorithm. In a total of 421 delinquent contracts, there were 116 contracts recovered. The model was able to correctly predict 106 contracts, reaching a recall of 94.38%. In addition, the model correctly predicted 297 (96.74%) contracts out of 307 contracts that would not be recovered.

Mortgage Loan III - The model with the highest recall was obtained with the GBM algorithm. In a total of 884 delinquent contracts, there were 212 contracts recovered. The model was able to accurately predict 186 contracts, reaching a recall of 98.94%. In addition, the model correctly predicted 670 (99.70%) contracts out of 672 contracts that would not be recovered.

The predictive models obtained from the analysis of the first three segments, out of a total of 11, have already shown a potential great benefit to the bank, effectively assisting its customers with delayed operations and avoiding unnecessary efforts in attempts in attempts of negotiation in contracts with low probability of recovering.

5.1 Future Works

The results obtained so far strengthen initiatives for the development of predictive models using machine learning techniques in the Bank studied.

With the increase in the efficiency of credit recovery, the Bank will benefit from the reduction in Allowance for Loan Losses (PDA), directly promoting positive results, with the reversal of provisions already made.

Thus, the study will be expanded to the 8 segments that have not yet been modeled, increasing the use of models obtained through machine learning techniques in credit recovery. In addition, the models al-

ready obtained can be improved with the use of ensemble models.

References

[1] Rogerio G. Lopes, Rommel N. Carvalho, Marcelo Ladeira, and Ricardo S. Carvalho. Predicting Recovery of Credit Operations on a Brazilian Bank. pages 780–784. IEEE, December 2016.

[2] Sung Ho Ha. Behavioral assessment of recoverable credit of retailer's customers. *Inf. Sci.*, 180(19):3703–3717, October 2010.

[3] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[4] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211 – 229, 2012.

[5] Karel Dejaeger, Frank Goethals, Antonio Giangreco, Lapo Mola, and Bart Baesens. Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, 218(2):548 – 562, 2012.

[6] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[7] Stephan Arndt, Carolyn Turvey, and Nancy C Andreasen. Correlating and predicting psychiatric symptom ratings: Spearmans r versus kendalls tau correlation. *Journal of psychiatric research*, 33(2):97–104, 1999.

[8] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.