RODRIGO KUSMINSKY HERSCU

Desenvolvimento e implantação de um modelo de detecção de fraude em cheques

São Paulo

2017

RODRIGO KUSMINSKY HERSCU

Desenvolvimento e implantação de um modelo de detecção de fraude em cheques

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Diploma de Engenheiro de Produção.

São Paulo

2017

RODRIGO KUSMINSKY HERSCU

Desenvolvimento e implantação de um modelo de detecção de fraude em cheques

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Diploma de Engenheiro de Produção.

Orientador:
Prof. Dr. Marco Aurélio de Mesquita

São Paulo

2017

Catalogação-na-publicação

Herscu, Rodrigo Kusminsky
Desenvolvimento e implantação de um modelo de detecção de fraude em cheques / R. K. Herscu -- São Paulo, 2017.
70 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1.detecção de fraudes 2.mineração de dados 3.aprendizagem de máquina 4.fraude financeira I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

À minha família.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Marco Aurélio de Mesquita, pela dedicação e paciência que possibilitaram a conclusão deste trabalho. Também agradeço a todos professores com quem tive contato ao longo da minha formação como engenheiro.

À minha família pelo apoio incondicional ao longo de toda minha trajetória. Sem eles nada disso seria possível.

Aos meus amigos da POLI e UNSW que tornaram os últimos seis anos muito melhores.

“In God we trust, all others bring data.”

(William Edwards Deming)

RESUMO

Este trabalho de formatura apresenta o desenvolvimento e a implantação de um modelo de detecção de fraudes em cheques em uma instituição financeira brasileira de grande porte, entre as cinco maiores do país. Após uma revisão bibliográfica sobre fraudes financeiras, conceitos de mineração de dados e algoritmos de aprendizagem de máquina, foi escolhida a metodologia CRISP-DM para guiar as atividades de desenvolvimento deste projeto. Após estudos sobre a performance de diferentes algoritmos e parâmetros sobre a seleção da amostra, foi realizado o treinamento do modelo utilizando o algoritmo *Random Forest* em uma base de cem mil linhas artificialmente balanceada para um nível de 50% de fraudes. A implementação do modelo incluiu a definição das atividades da área e levou a uma redução das perdas com fraude e do custo operacional da área de conferência de cheques.

Palavras-chave: detecção de fraudes, mineração de dados, aprendizagem de máquina, fraude financeira

ABSTRACT

This paper presents the development and deployment of a cheque fraud detection model in a large Brazilian financial institution, among the five largest in the country. Following a literature review regarding financial frauds, data mining concepts and machine learning algorithms, the CRISP-DM methodology was chosen do guide the activities of this project development. After analyzing the performance of different algorithms and parametrizations, a hundred thousand lines, artificially balanced to a 50% fraud index, database was used in the training with the Random Forest algorithm. The model deployment included the definition of the division activities and led to a decrease in both financial fraud loss and operational costs due to cheque clearing activities.

Key words: fraud detection, data mining, machine learning, financial fraud

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AUC	*Area Under Curve*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DMAIC	*Define, Measure, Analyze, Improve and Control*
KDD	*Knowledge Discovery in Databases*
PDCA	*Plan, Do, Check, Act*
ROC	*Receiver Operating Characteristic*
SEMMA	*Sample, Explore, Modify, Model & Assess*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta a contextualização e relevância do trabalho, assim como os objetivos propostos e a estruturação dos seus capítulos.

1.1 Contexto e relevância

1.1.1 Sobre o Estágio

O trabalho aqui apresentado foi desenvolvido pelo autor durante o estágio acadêmico realizado em uma empresa de consultoria de gestão. O trabalho teve como cliente um grande banco de varejo, que está entre os cinco maiores do país.

O projeto foi desenvolvido entre março e outubro de 2017 em uma equipe de três pessoas. As principais atividades do autor foram: a realização de entrevistas para entendimento dos processos que envolvem a compensação de cheques, levantamento dos tipos de fraude e de hipóteses sobre variáveis que possam explicar a fraude, coleta e mapeamento de bases de dados, cálculo de variáveis, análise exploratória de dados, modelagem matemática, implantação e definição dos processos de rotina de aplicação do modelo.

1.1.2 Cheques

O cheque é uma ordem de pagamento à vista e um título de crédito. A operação padrão de um cheque envolve três agentes: o emitente, o beneficiário e o sacado. O primeiro é aquele que emite o cheque, o segundo é quem será favorecido com o montante descrito no documento e o terceiro é o banco no qual está o depositado dinheiro do emitente (Banco Central do Brasil, 2014).

O beneficiário pode receber o dinheiro de duas formas, a primeira delas é retirando o dinheiro em espécie diretamente no banco sacado. A segunda forma é através do processo de compensação de cheques. Nesse caso o valor pode ser depositado diretamente em sua conta, podendo esta estar vinculada a qualquer instituição financeira vinculada ao sistema nacional de compensação. É importante notar que as instituições financeiras são obrigadas por lei a oferecer essa modalidade de forma gratuita, pois é considerada um serviço essencial (Banco Central do Brasil, 2014).

A compensação de cheques deve ser efetuada através de imagem digital. Os cheques depositados ao longo do dia têm sua imagem capturada ao final do expediente, e são enviados ao Banco do Brasil, que por sua vez redistribui as imagens para os bancos sacados, para que seja aprovado o pagamento. O pagamento será aprovado mediante fundos do emitente e

verificação de questões formais referentes ao documento, como autenticidade da folha de cheque, assinatura do emitente e preenchimento correto conforme determinam as regras. Em caso de alguma irregularidade o cheque será devolvido, e dependendo do motivo de devolução poderá ou não ser reapresentado.

O uso dos cheques vem caindo no Brasil como pode ser observado na Figura 1, caindo ano a ano e atingindo um patamar mínimo de 879 milhões de transações no ano de 2016. Entretanto, ao observarmos o valor transacionado (Figura 2), observamos que este meio de pagamento ainda é extremamente relevante no país, com um valor total transacionado de aproximadamente 2,26 trilhões de reais no ano de 2016, mais do que o dobro do valor transacionado através de cartões de crédito e débito juntos. Os dados são das Estatísticas de Pagamentos de Varejo e Cartões no Brasil (2016), divulgadas anualmente pelo Banco Central do Brasil, disponíveis na Tabela 1.

Tabela 1 – Dados das Estatísticas de Pagamentos de Varejo e Cartões no Brasil 2016

	Quantidade de Transações (milhões)			**Valor das Transações (R$ bilhões)**		
Ano	**Cheque**	**Cartão de Débito**	**Cartão de Crédito**	**Cheque**	**Cartão de Débito**	**Cartão de Crédito**
2008	1.963	2.136	2.546	2.554	107	220
2009	1.803	2.445	2.808	2.502	128	263
2010	1.675	2.948	3.314	2.691	159	332
2011	1.590	3.508	3.836	2.786	196	401
2012	1.439	4.129	4.473	2.891	237	465
2013	1.297	4.908	5.020	2.917	293	534
2014	1.165	5.627	5.367	2.801	348	594
2015	1.018	6.467	5.559	2.579	390	653
2016	879	6.837	5.858	2.259	430	674

(Fonte: Banco Central do Brasil)

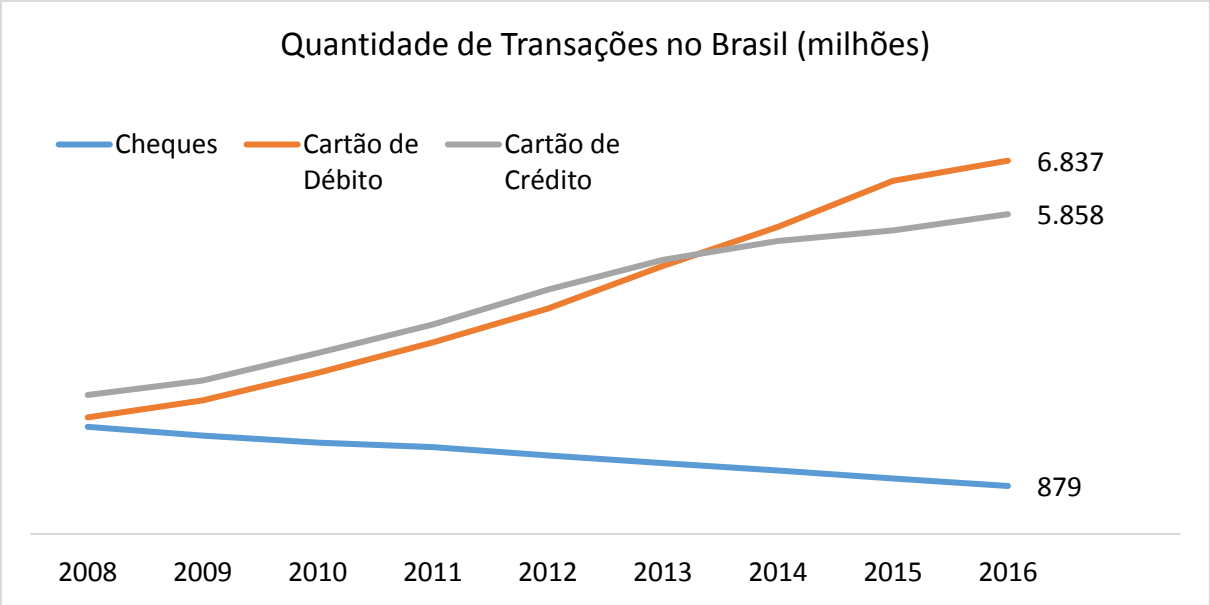

Figura 1 – Evolução da quantidade de transações através dos principais meios de pagamento no Brasil

(Fonte: Banco Central do Brasil)

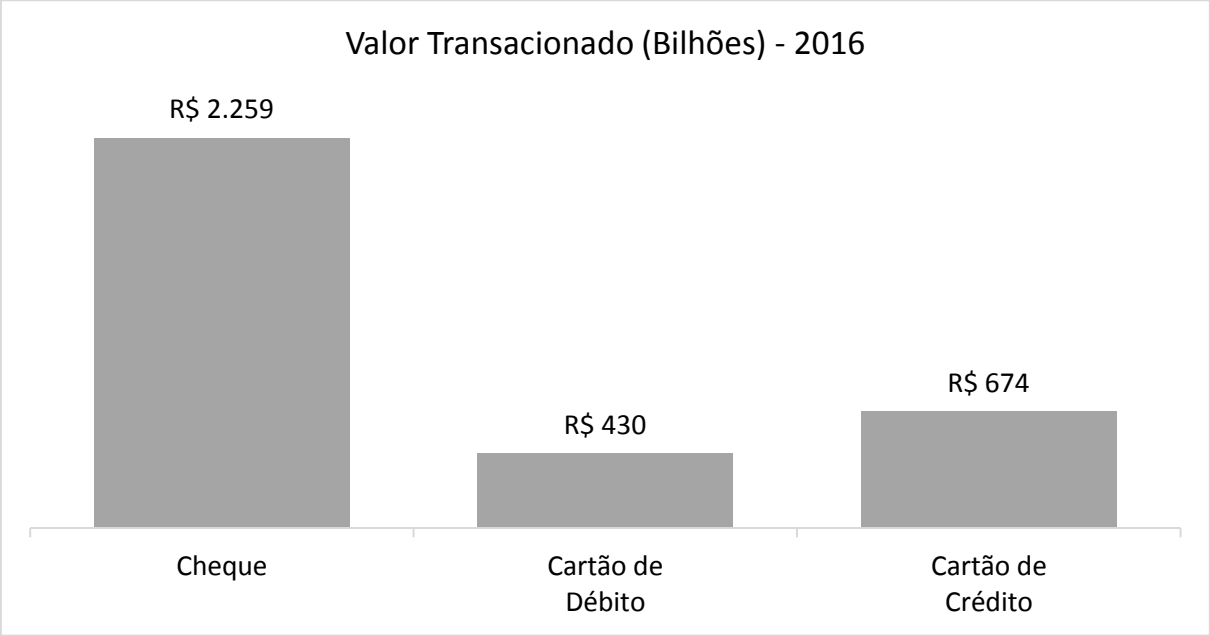

Figura 2 – Valor transacionado através dos principais meios de pagamento no Brasil em 2016

(Fonte: Banco Central do Brasil)

Classificação	Motivo	Descrição
I - Cheque sem provisão de fundos	11	Cheque sem fundos - 1ª apresentação
	12	Cheque sem fundos - 2ª apresentação
	13	Conta encerrada
	14	Prática espúria
II - Impedimento ao pagamento	20	Cheque sustado ou revogado em virtude de roubo, furto ou extravio de folhas de cheque em branco
	21	Cheque sustado ou revogado
	22	Divergência ou insuficiência de assinatura
	23	Cheques emitidos por entidades e órgãos da administração pública federal direta e indireta, em desacordo com os requisitos constantes do art. 74, § 2º, do Decreto-Lei nº 200, de 25.2.1967
	24	Bloqueio judicial ou determinação do Banco Central do Brasil
	25	Cancelamento de talonário pelo participante destinatário
	26	Inoperância temporária de transporte
	27	Feriado municipal não previsto
	28	Cheque sustado ou revogado em virtude de roubo, furto ou extravio
	30	Furto ou roubo de cheque
	70	Sustação ou revogação provisória
III - Cheque com irregularidade	31	Erro formal (sem data de emissão, com o mês grafado numericamente, ausência de assinatura ou não registro do valor por extenso)
	33	Divergência de endosso
	34	Cheque apresentado por participante que não o indicado no cruzamento em preto, sem o endosso-mandato
	35	Cheque fraudado, emitido sem prévio controle ou responsabilidade do participante ("cheque universal"), ou ainda com adulteração da praça sacada, ou ainda com rasura no preenchimento
IV - Apresentação indevida	37	Registro inconsistente
	38	Assinatura digital ausente ou inválida
	39	Imagem fora do padrão
	40	Moeda inválida
	41	Cheque apresentado a participante que não o destinatário
	42	Cheque não compensável na sessão ou sistema de compensação em que apresentado
	43	Cheque, devolvido anteriormente pelos motivos 21, 22, 23, 24, 31 e 34, não passível de reapresentação em virtude de persistir o motivo da devolução
	44	Cheque prescrito
	45	Cheque emitido por entidade obrigada a realizar movimentação e utilização de recursos financeiros do Tesouro Nacional mediante Ordem Bancária
	48	Cheque de valor superior a R$ 100,00 (cem reais), emitido sem a identificação do beneficiário
	49	Remessa nula, caracterizada pela reapresentação de cheque devolvido pelos motivos 12, 13, 14, 20, 25, 28, 30, 35, 43, 44 e 45
V - Emissão indevida	59	Informação essencial faltante ou inconsistente não passível de verificação pelo participante remetente e não enquadrada no motivo 31
	60	Instrumento inadequado para a finalidade
	61	Item não compensável
	64	Arquivo lógico não processado / processado parcialmente
VI - A serem empregados diretamente pela instituição financeira contratada	71	Inadimplemento contratual da cooperativa de crédito no acordo de compensação
	72	Contrato de compensação encerrado

Figura 3- Motivos de Devolução de Cheques

(Fonte: Banco Central do Brasil)

1.1.3 Fraudes

As fraudes em cheques podem ter como vítima o beneficiário, isso ocorre quando ao receber um cheque como meio de pagamento, fica impossibilitado de receber por falta de fundos, revogação do cheque, entre outros.

A vítima também pode ser o emitente, isso ocorre quando o fraudador se apropria de uma folha original ou a falsifica e utiliza o cheque para receber fundos de forma indevida ou trocá-lo por mercadorias e serviços. Nesse caso, comprovada a fraude, o banco sacado deve ressarcir o emitente, sendo a instituição financeira quem arca com as perdas da fraude nesse tipo de

cenário. Estima-se que em 2010 os bancos atuantes no Brasil perderam aproximadamente R$ 1,25 bilhões de reais em fraudes envolvendo cheques (OLIVEIRA, 2012).

O processo compensação eletrônica visa reduzir ao máximo essas perdas. Nesse processo, operadores verificam as imagens capturadas dos cheques e conferem a validade da assinatura, autenticidade da folha e outras questões formais sobre o preenchimento do cheque. Entretanto, a atuação especializada de fraudadores, torna cada vez mais difícil a identificação dessas irregularidades.

1.2 Problema

Pelo volume da operação, na casa de centenas de milhares de cheques por dia, são necessárias centenas de operadores, que possuem níveis de capacitação diferentes. Dessa forma, atribuir os cheques de maior risco para os melhores operadores pode garantir uma significativa redução das perdas financeiras com fraudes em cheques compensados por parte dos bancos. Além disso, ao priorizar a conferência dos cheques mais críticos é possível reduzir o tempo total gasto pela instituição financeira na conferência de cheques, gerando ganhos em eficiência operacional.

Neste contexto surgem os objetivos deste trabalho.

1.3 Objetivos

O presente trabalho tem por objetivo criar um modelo preditivo que indique, diariamente, a propensão a ser uma fraude que cada cheque recebido no dia pelo banco possui. A partir de tal modelo será feita a seleção de quais cheques apresentam maior risco para a instituição financeira, e devem ter sua conferência priorizada.

A detecção de fraudes financeiras, por envolver a extração e descoberta de informações ocultas em grandes quantidades de dados é um típico problema de *Data Mining* (NGAI, HU, WONG, CHEN, & SUN, 2010).

Phua et al. (2005) ainda apontam que a detecção de fraudes se tornou uma das aplicações mais bem estabelecidas para as técnicas de *Data Mining*, tanto na indústria quanto no governo.

1.4 Estrutura do trabalho

A estrutura deste trabalho é dividida em seis capítulos, sendo esta introdução o primeiro deles, abordando o contexto e a relevância do tema estudado, a definição do problema a ser resolvido e os objetivos propostos.

O segundo capítulo apresenta a revisão da literatura relevante ao tema estudado, apresentando os principais conceitos relativos a fraudes financeiras, um estudo sobre o tema de mineração de dados e aprendizagem de máquina, além de uma revisão sobre metodologias de resolução de problemas.

No Capítulo 3 as metodologias revisadas no capítulo anterior são comparadas para que seja definida a que será utilizada no desenvolvimento do trabalho.

Os Capítulos 4 e 5 são dedicados ao desenvolvimento do projeto, que foi dividido em duas fases, sendo a segunda referente a implantação do que foi desenvolvido na Fase I.

O último capítulo apresenta a conclusão do projeto, uma breve discussão dos resultados alcançados, desafios enfrentados e possíveis próximos passos e desdobramentos deste trabalho.

2 REVISÃO BIBLIOGRÁFICA

Esse capítulo é dedicado ao estudo dos principais temas relevantes para o desenvolvimento deste trabalho.

2.1 Fraudes

Fraude é um ato ilícito que tem como objetivo a obtenção de vantagens indevidas, normalmente de forma prejudicial a terceiros. Existem indícios da prática de atos fraudulentos que remontam a origem da economia, sendo, portanto, um problema antigo (MARCONDES, 2017).

Marcondes (2017) aponta que a incidência de fraudes pode ser explicada pela ocorrência de três fatores simultâneos:

- A existência de golpistas motivados,
- A disponibilidade de vítimas vulneráveis,
- Regras, normas e meios de controle pouco eficazes ou inexistentes.

De forma geral, o ato fraudulento pode ser constituído na sequência de etapas em que o fraudador constituiu uma falsa representação da realidade, uma vítima aceita tal representação como verdadeira, de forma que ela é prejudicada em benefício do fraudador, comumente de forma financeira (GOLDEN, SALAK, & CLAYTON, 2006).

2.1.1 Fraude Financeira

De acordo com a *Cornell University Law School* (2009), fraude financeira é a execução consciente, ou tentativa de, de esquema ou artifício contra uma instituição financeira para a obtenção e dinheiro, fundos, crédito, ativos, seguros, etc. de propriedade da instituição financeira, ou sob custódia da mesma.

A fraude financeira é um problema de amplas consequências, pois atinge não só a instituição financeira como a vida cotidiana de quem interage com ela, como clientes e seus parceiros comerciais. De forma geral, a incidência de fraudes reduz a confiança no mercado, desestabiliza a economia e afeta o custo da instituição e por consequência o custo de vida das pessoas (NGAI, HU, WONG, CHEN, & SUN, 2010).

Com relação a cheques, é considerada fraude a transação que não partiu do emitente, ou seja, a emissão de um cheque de uma conta sem o conhecimento ou autorização de seu titular. Isso pode acontecer de duas formas. Uma delas é quando o fraudador clona a folha de cheque, a outra é quando ele obtém uma folha original e a utiliza fingindo ser o real titular da conta. Em

ambos os casos, o fraudador ainda terá que forjar a assinatura, que será verificada pela instituição financeira. Em ambas as modalidades, o banco é considerado responsável e deverá ressarcir o cliente, como explica a advogada Maria Elisa Novais, em reportagem jornalística (ISTOÉ, 2016).

2.1.2 Detecção de Fraudes Financeiras

Tradicionalmente, a detecção de fraudes financeiras dependia principalmente de técnicas manuais, que são ineficientes e podem ser pouco confiáveis, dada a dificuldade de certos problemas. Atualmente, abordagens baseadas em *Data Mining* vêm se mostrando promissoras, dada sua capacidade de identificar pequenas anomalias em grandes volumes de dados (NGAI, HU, WONG, CHEN, & SUN, 2010).

West & Bhattacharya (2014) argumentam que os métodos manuais além de caros, imprecisos e consomem muito tempo, são impraticáveis na era do *Big Data*. O objetivo da detecção de fraudes financeiras é, a partir de uma grande quantidade de dados transacionais, apontar a transações fraudulentas em meio às legítimas e, portanto, se trata de um tradicional problema de *Data Mining* (WEST & BHATTACHARYA, 2016).

Outra vantagem deste tipo de abordagem reside no fato de que fraudadores refinam seus métodos de forma contínua, e é necessário um método de detecção que também evolua, e seja capaz de identificar novos padrões, como é o caso das técnicas de *Machine Learning* (aprendizagem de máquina), largamente utilizados em problemas de *Data Mining* (WEST & BHATTACHARYA, 2014).

A detecção de fraudes é considerada um problema de classificação, pois seu objetivo é categorizar as transações em um de dois grupos: legítimas ou fraudulentas. Uma particularidade deste tipo de problema está no desbalanceamento das transações, pois tradicionalmente o volume de transações legítimas é muito maior do que o de transações fraudulentas, o que exige tratamentos de dados específicos para que sejam atingidos bons resultados (WEST & BHATTACHARYA, 2014).

2.2 Data Mining e exemplos de aplicação

Data Mining pode ser definido como um processo que utiliza técnicas estatísticas, matemáticas, de Inteligência Artificial e de *Machine Learning* para extrair e identificar informações úteis para a tomada de decisão a partir de um grande volume de dados (TURBAN, ARONSON, LIANG, & SHARDA, 2007).

West & Bhattacharya (2014) definem *Data Mining* como qualquer metodologia que processe grandes quantidades de dados a fim de revelar significados ocultos.

Machine Learning é uma subárea da Inteligência Artificial, também conhecida como aprendizado de máquina. É baseada em algoritmos com capacidade de aprender determinado padrão ou comportamento a partir de observações prévias. Dessa forma, possui como objetivo maior a predição de características conhecidas a partir dessas observações, chamadas de dados de treinamento.

Os termos *Data Mining* e *Machine Learning* são comumente vistos conjuntamente e eventualmente confundidos, portanto é importante esclarecer suas diferenças. Enquanto *Data Mining* possui o significado mais amplo de buscar propriedades desconhecidas nos dados, *Machine Learning* são algoritmos que são capazes de aprender com dados históricos, e, portanto, largamente empregados em problemas de *Data Mining*.

Em uma conferência da empresa de software SAS em 1998, foi apresentado um diagrama que buscava apresentar as diferenças e intersecções entre essas áreas do conhecimento, bem como outras áreas relevantes. Nota-se que diversas técnicas e assuntos se correlacionam com *Data Mining*, que as utiliza como meio para alcançar seus objetivos. Na Figura 4, *Data Mining* é apresentado dentro da área de conhecimento chamada de KDD (*Knowledge Discovery in Databases*), da qual é a principal área de estudos.

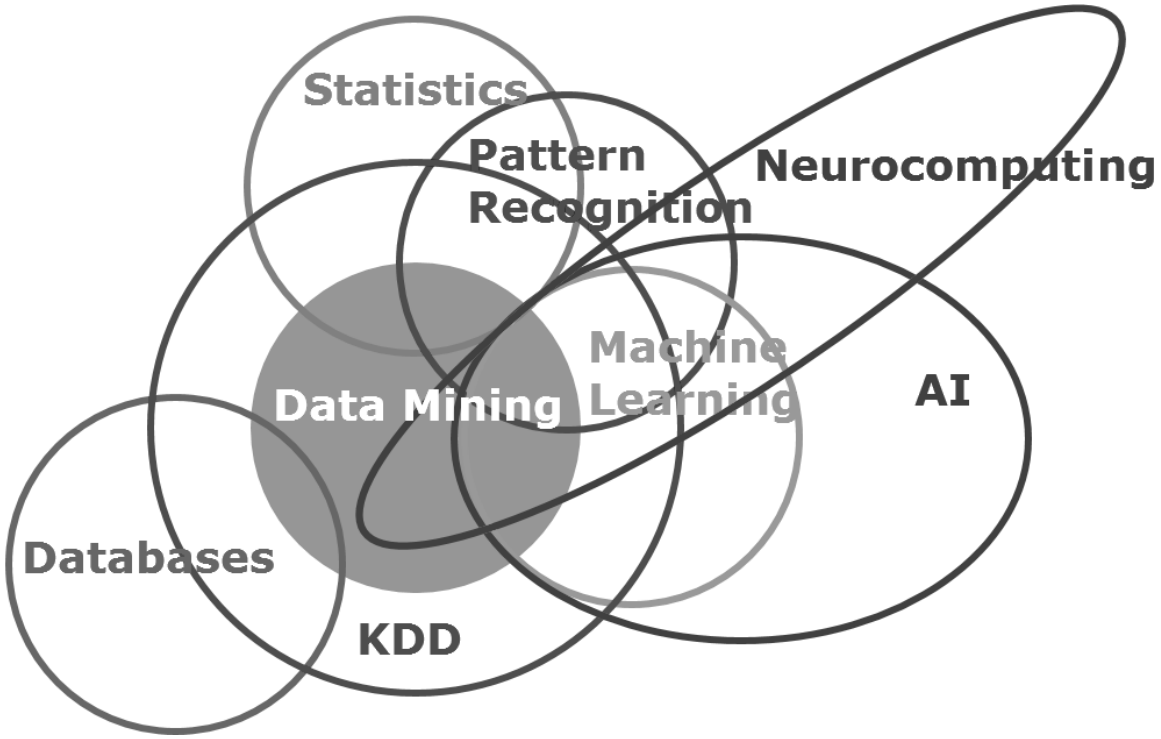

Figura 4 - Diagrama apresentado na primeira conferência do SAS sobre Data Mining, em 1998 (Fonte: SAS Institute)

As próximas seções tratarão sobre os algoritmos e técnicas de *Machine Learning* estudados para utilização neste trabalho, uma discussão sobre o conceito de *overfitting* e seus impactos e soluções, uma análise sobre métricas de performance utilizadas na avaliação de modelos, e uma discussão sobre a questão do balanceamento dos dados, típica em problemas de detecção de fraude.

2.2.1 Regressão Logística

A Regressão Logística é um modelo linear generalizado que utiliza a distribuição binomial como função de ligação. Ela é comumente utilizada em cenários em que a variável resposta seja binária. É possível utilizar a metodologia para estimar qual a probabilidade de que cada item pertença a determinada classe, sendo amplamente utilizada na detecção de fraudes na indústria financeira (PROVOST & FAWCETT, 2013).

Para garantir que a soma das probabilidades de pertencimento às classes resulte exatamente em 100% é utilizado o chamado operador *logit* (Equação 1). Visualmente, fica claro como esse operador desloca os resultados para o intervalo desejado, como pode ser observado na Figura 5.

$$p(x) = \frac{1}{1 + e^{-f(x)}} \quad (1)$$

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (2)$$

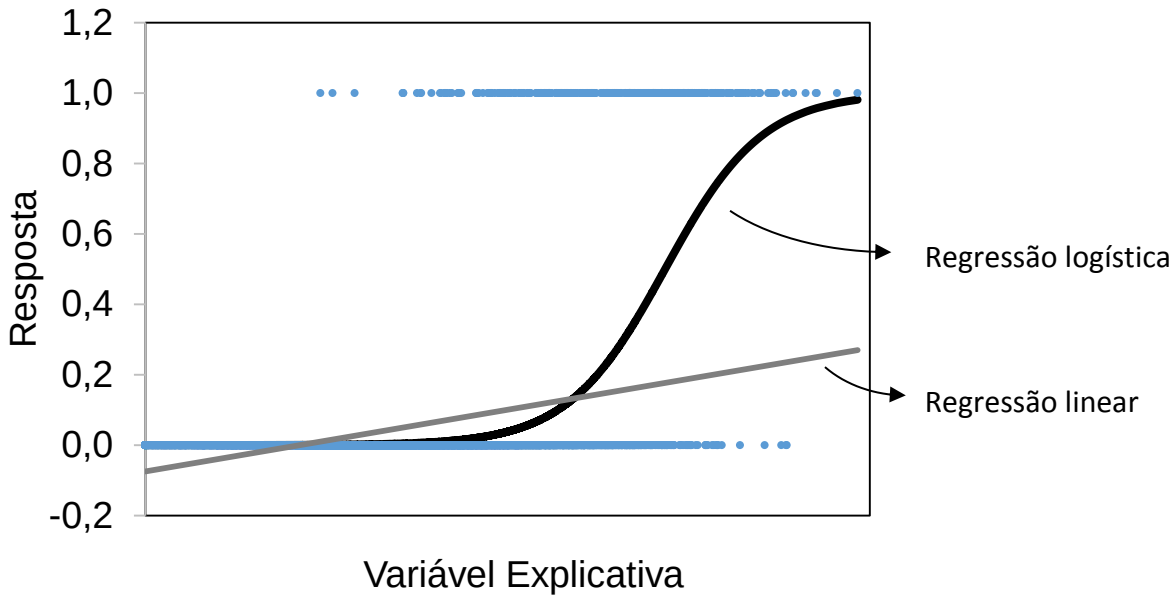

Figura 5 - Exemplo de como o operador *logit* desloca a regressão linear para o intervalo desejado.

(Fonte: elaborado pelo autor)

A modelagem através do método de Regressão Logística consiste em, a partir dos dados históricos, calcular os parâmetros beta (Equação 2) que minimizem os quadrados dos erros ou, alternativamente, que minimizem os erros de classificação, ou seja, apontar que um item pertence a determinada classe quando na verdade não pertence.

A Equação 2 pode ser aplicada a novos itens cuja classe seja desconhecida. Dessa forma, é obtida uma estimativa da probabilidade de que aquele item pertença ou não à determinada classe (HASTIE, TIBSHIRANI, & FRIEDMAN, 2001).

2.2.2 Árvore de Decisão

A técnica de árvores de decisão consiste em classificar dados utilizando uma estrutura de árvore, na qual os nós representam escolhas binárias baseadas nos atributos de cada item, os nós dividem os dados em subgrupos mutuamente exclusivos, o que pode ser feita de forma recursiva até um determinado ponto de parada (WEST & BHATTACHARYA, 2014).

Diversas estratégias podem ser utilizadas para gerar uma árvore de decisão, sendo a mais comum delas baseado no ganho de informação. Nesse método, a partir de dados previamente

classificados, é escolhido o atributo que melhor divide esses dados em subgrupos, sendo o critério para definição da melhor divisão aquela que gera o maior ganho de informação (PROVOST & FAWCETT, 2013).

O ganho de informação é definido como a diferença entre a entropia de um nó com a média ponderada pela probabilidade da entropia de seus subgrupos (Equação 4), sendo a entropia (Equação 3) uma medida da desorganização dos dados (PROVOST & FAWCETT, 2013).

$$entropia = -p_1 \log(p_1) - p_2 \log(p_2) - \cdots - p_n \log(p_n) \tag{3}$$

$$IG = entropia(C) - [p(c_1)entropia(c_1) + p(c_2)entropia(c_2) + \cdots + p(c_n)entropia(c_n)] \tag{4}$$

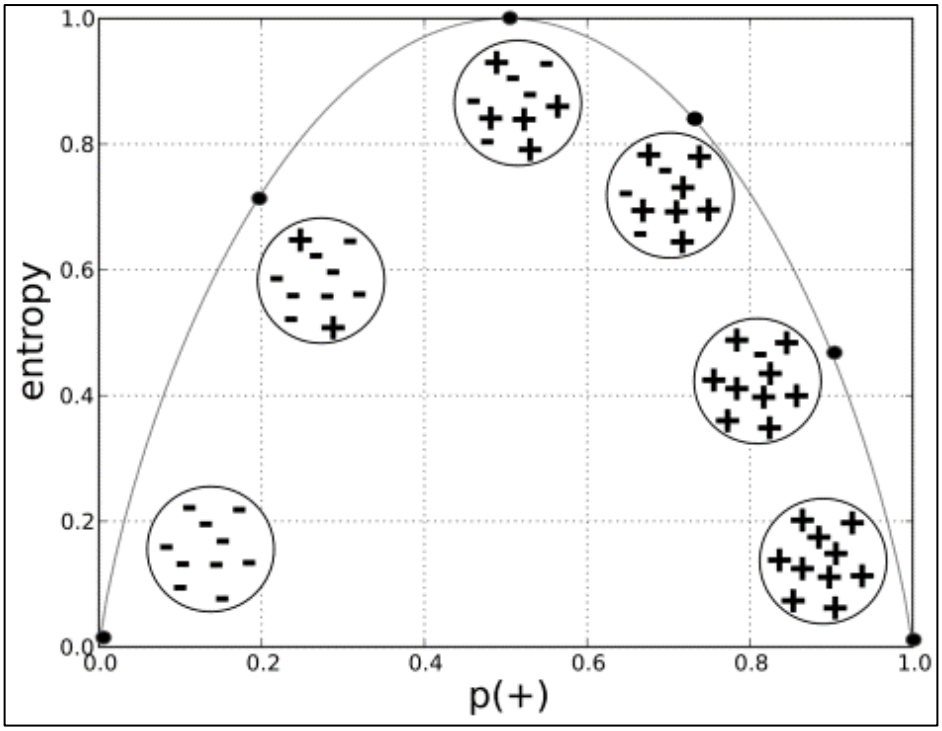

Figura 6 - Variação da entropia em função da pureza de um grupo, exemplo de classificação binária.

(Fonte: Provost & Fawcett)

Provost & Fawcett (2013) apontam que embora existam modelos mais precisos, as árvores de decisão têm alto potencial explicativo e são especialmente vantajosas pela facilidade em interpretá-las, sendo facilmente explicada para e aceita por Stakeholders que não tenham conhecimento prévio acerca de *Data Mining* e algoritmos mais complexos.

2.2.3 Random Forests

Bhattacharyya, Jha, Tharakunnel & Westland (2011) alertam que modelos que utilizam uma única árvore de decisão podem ser instáveis e excessivamente sensíveis a dados específicos utilizados em sua construção. Esse tipo de problema pode ser endereçado por modelos compostos, isso é, quando se agrega uma série de modelos para chegar a uma resposta.

Random Forests são uma categoria de modelos compostos na qual se criam diversas árvores de decisão para o mesmo problema, e se utilizam todas elas de forma conjunta para se chegar em uma resposta final. As diferentes árvores são construídas utilizando-se de duas fontes de aleatoriedade: primeiramente, cada árvore é construída com base em apenas uma amostra dos dados e além disso só é considerado um subconjunto das variaveis explicativas disponíveis na construção de cada árvore. As previsões decorrentes desse tipo de modelo são obtidas pela agregação do resultado que cada árvore apresentou individualmente. Para casos de classificação, pode-se utilizar como escolha final, aquela escolhida pela maioria das árvores, como numa votação (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

Bhattacharyya, Jha, Tharakunnel & Westland (2011) defendem que *Random Forests* são computacionalmente eficientes, pois cada árvore é construída de forma independente das demais. Além dissom indicam que ao atingir um certo número de árvores, o modelo torna-se robusto contra problemas típicos das árvores de decisão individuais, como *overfitting* e ruídos específicos dos dados utilizados na construção do modelo.

2.2.4 Overfitting

As técnicas utilizadas em *Machine Learning* buscam encontrar padrões nos dados que sirvam para responder a perguntas adicionais sobre novos dados sobre os quais ainda não sabemos a resposta. O fenômeno de identificarmos padrões nos nossos dados de treinamento, porém eles não poderem ser generalizados para novas instâncias é chamado de *overfitting* (PROVOST & FAWCETT, 2013).

Provost & Fawcett (2013) definem *overfitting* como a tendência dos procedimentos de *Data Mining* de produzir modelos excessivamente adaptados à base de treinamento, ao custo da generalização a novos dados. Segundo eles, conforme a complexidade do modelo aumenta, maior sua liberdade para apontar correlações espúrias, específicas dos dados utilizados no treinamento, e que não representam características da população em geral. A Figura 7 apresenta a variação da assertividade de um modelo particular em função de sua complexidade. A complexidade nesse caso é representada pelo número máximo de nós que é permitido a um

modelo de árvore de decisão. É possível observar que o aumento de complexidade sempre leva a uma maior assertividade quando esta é medida nos dados utilizados para o treinamento do modelo. Entretanto ao ser medida em dados de teste, a partir de certo ponto passa a ocorrer uma degradação dessa assertividade.

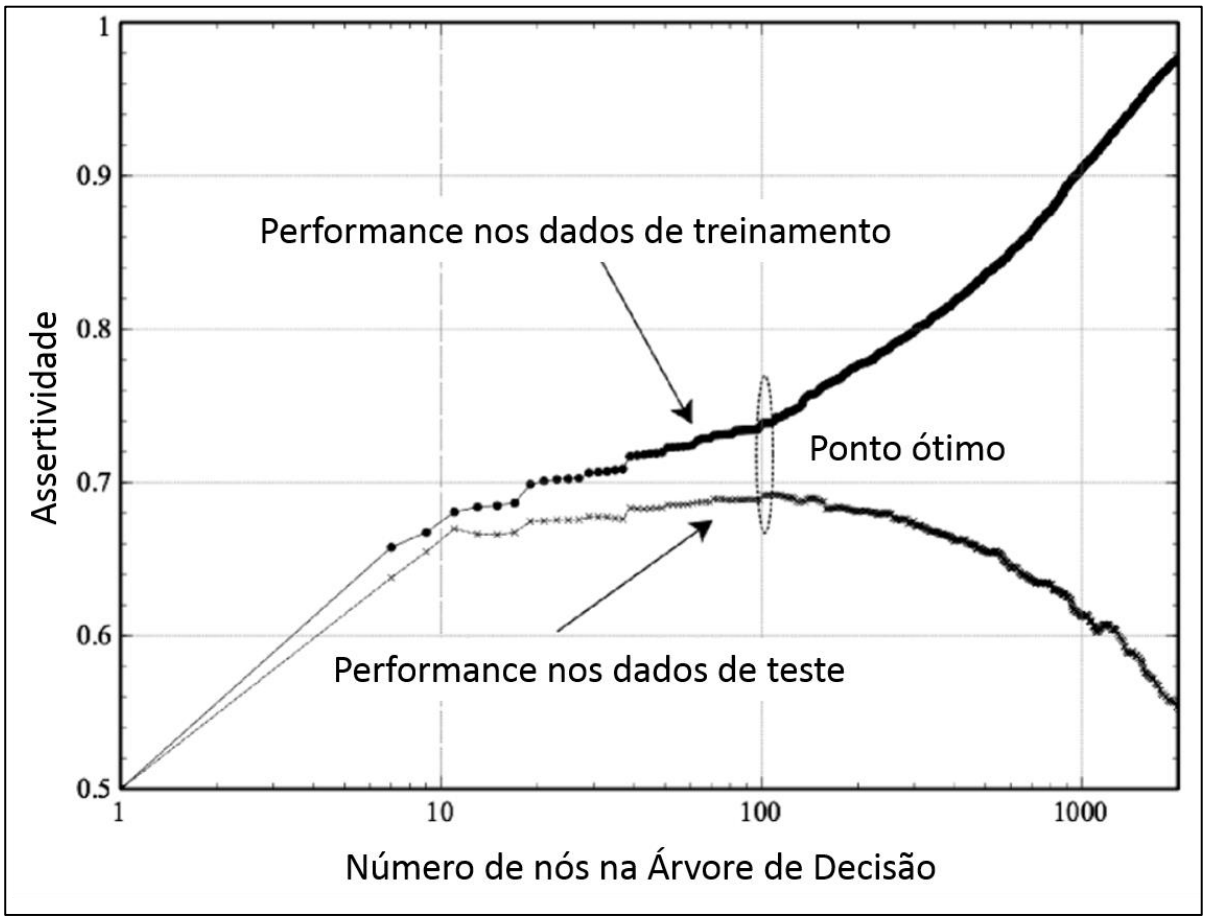

Figura 7 - Assertividade em função da complexidade para dados de treino e teste. (Fonte: Provost & Fawcett)

Dessa forma, a avaliação de um modelo a partir dos dados utilizados em seu treinamento não pode ser generalizada para como ele irá se comportar em novos dados. Dessa forma surge a necessidade de resguardar uma parcela dos dados para serem utilizadas apenas na avaliação do modelo. Em resumo, a partir de um volume de dados cuja resposta em conhecida, parte dele será utilizada para construção do modelo, chamada de base de treinamento. Esse modelo será aplicado à parcela restante dos dados, e o resultado será comparado com a resposta real que é conhecida, permitindo identificar a qualidade do modelo. A essa parcela restante damos o nome de base de testes. A separação dos dados entre base de treinamento e de testes é feita de forma aleatória (PROVOST & FAWCETT, 2013).

2.2.5 Métricas

A capacidade de generalização de um modelo de *Machine Learning* está relacionada com sua capacidade de predizer corretamente quando aplicado em novas séries de dados. No caso de detecção de fraudes, o problema pode ser interpretado como uma classificação binária, e predizer corretamente consiste em classificar corretamente transações em fraudulentas ou não fraudulentas. Medir esta capacidade de generalização é importante, pois guia a escolha de qual deve ser o modelo escolhido e seus parâmetros, além de ser uma medida da qualidade do modelo escolhido (HASTIE, TIBSHIRANI, & FRIEDMAN, 2001).

Avaliar um modelo de classificação binária consiste em comparar a quantidade de classificações corretas com as incorretas geradas pelo modelo, isso pode ser feito a partir de uma matriz de confusão como exemplificado na Figura 8.

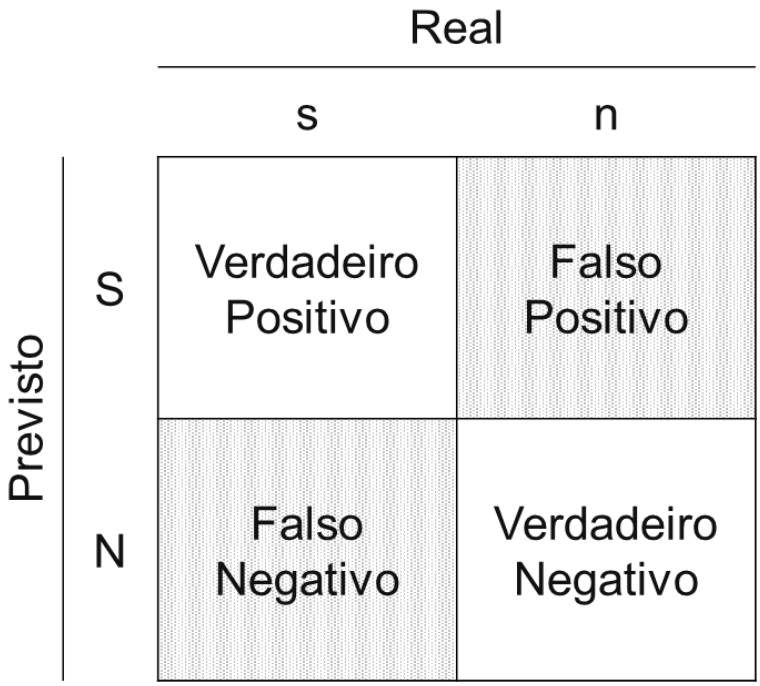

Figura 8 - Matriz de confusão
(Fonte: elaborado pelo autor)

Métricas comuns na avalição de modelos são a acurácia (Equação 5), que mede a proporção de casos previstos corretamente, a sensibilidade (Equação 6), que mede a assertividade sobre os casos positivos, e a especificidade (Equação 7), que mede a assertividade sobre os casos negativos. As equações as seguir mostram suas relações com a matriz de confusão (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

$$Acurácia = \frac{Previsões\ corretas}{Total\ de\ previsões} = \frac{VP + VN}{VP + VN + FP + FN} \quad (5)$$

$$Sensibilidade = \frac{VP}{VP + FN} \quad (6)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (7)$$

A matriz de confusão, e por consequência as métricas acima, dependem de um valor estabelecido que serve como limite para sua classificação como positivo ou negativo. Esse valor geralmente é 0,5, pois se assume que casos com probabilidade acima de 50% de serem positivos são classificados como tal, porém não necessariamente. Uma alternativa para visualizar a qualidade do modelo como um todo são as curvas ROC (*Receiver Operating Characteristic*), que mostram a como a sensibilidade e especificidade variam com relação a essa escolha. Nota-se que cada ponto da curva representa uma matriz de confusão diferente, como mostra a Figura 9 (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

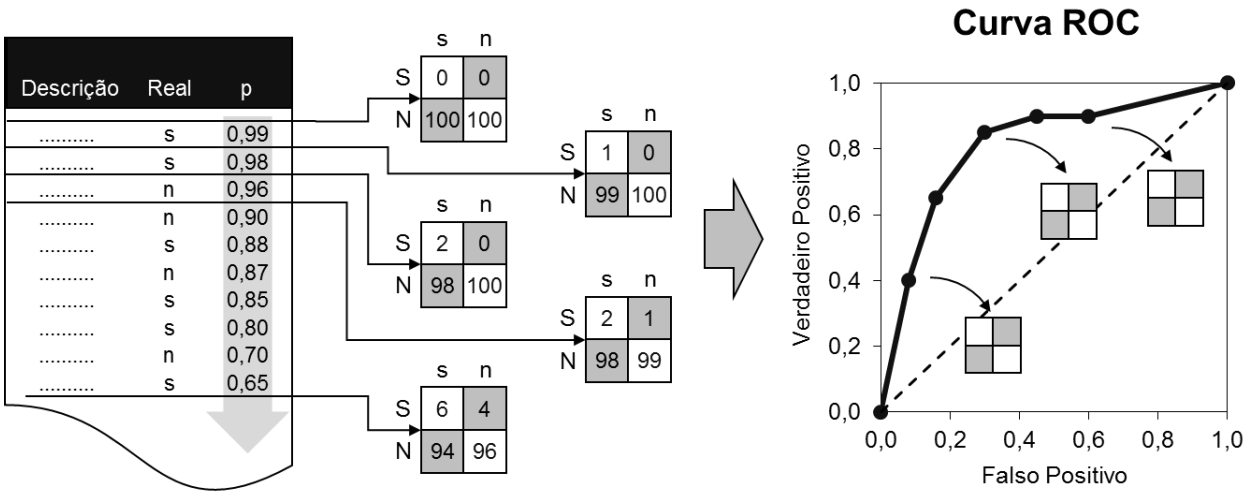

Figura 9 - Construção da curva ROC

(Fonte: adaptado de Provost & Fawcett)

A linha diagonal da curva ROC representa um modelo de decisão aleatório, portanto um modelo é tão bom quanto mais sua curva estiver distante da diagonal. A partir dela, pode ser obtida a métrica AUC (*Area Under Curve)*, sendo a área sob a curva ROC, que pode variar entre 0 e 1. É importante notar que um modelo aleatório teria AUC de 0,5.

Bhattacharyya, Jha, Tharakunnel, & Westland (2011) indicam que para casos desbalanceados a Acurácia não é uma boa métrica pois um modelo que sempre indicasse a classe predominante

se sairia excessivamente bem. Pozzolo, Caelen, Borgne, Waterschoot, & Bontempi (2014) afirmam que AUC e a curva ROC são boas métricas para casos onde há desbaleanceamento.

2.2.6 Questão do Balanceamento

Um desafio comum em problemas de detecção de fraudes está no desbalanceamento do problema, isso significa que uma das classes de um problema de classificação binário possui muito mais representantes do que a outra. No caso de detecção de fraudes, transações legítimas são muito mais comuns que as fraudulentas (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

Segundo Pozzolo, Caelen, Borgne, Waterschoot, & Bontempi (2014), a aprendizagem a partir de dados desbalanceados é uma tarefa complicada para a maioria dos algoritmos, pois eles não foram desenvolvidos para lidar com essa grande diferença entre a quantidade de membros de cada classe.

Este tipo de problema pode ser sanado por técnicas de amostragem que balanceiem os dados. Essas técnicas podem ser divididas entre aquelas que atuam no nivel dos dados e no nível algorítmico. As técnicas que atuam no nível algorítmico são mais complexas e estudos indicam que apresentam pior performance em casos de detecção de fraude. O método mais simples e de melhor performance consiste em realizar um *downsampling* da classe mais populosa, para reduzir sua proporção. Isso ocorre como uma etapa de pré-processamento dos dados, anterior a aplicação de qualquer algoritmo (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

2.3 Metodologias de Resolução de Problemas

Provost & Fawcett (2013) propõem que a extração de conhecimento dos dados para a resolução de problemas de negócio deve ser tratada de forma sistemática, seguindo um processo com etapas bem definidas. As metodologias mais utilizadas em problemas de mineração dados são conhecidas como CRISP-DM e SEMMA e guardam semelhanças com metodologias de solução de problemas clássicas como o ciclo PDCA de Deming e DMAIC. Nessa seção será apresentado um estudo das quatro metodologias, que serão comparadas no Capítulo 3 para que seja escolhida aquela que será utilizada neste trabalho.

2.3.1 CRISP-DM

A metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) foi criada em 1996 por um consórcio de empresas que estavam na vanguarda da aplicação de *Data Mining* a

negócios, com o objetivo de ser um modelo de referência passível de ser aplicado em qualquer indústria.

O modelo é representado pela Figura 10, na qual se observa que não se trata de uma sequência de fases fixas, sendo passível de idas e vindas, que são inclusive encorajadas. As setas indicam as relações e dependências mais frequentes entre as fases, e não uma sequência rígida de etapas. Além disso, é recomendado que o processo tenha natureza cíclica, de melhoria contínua.

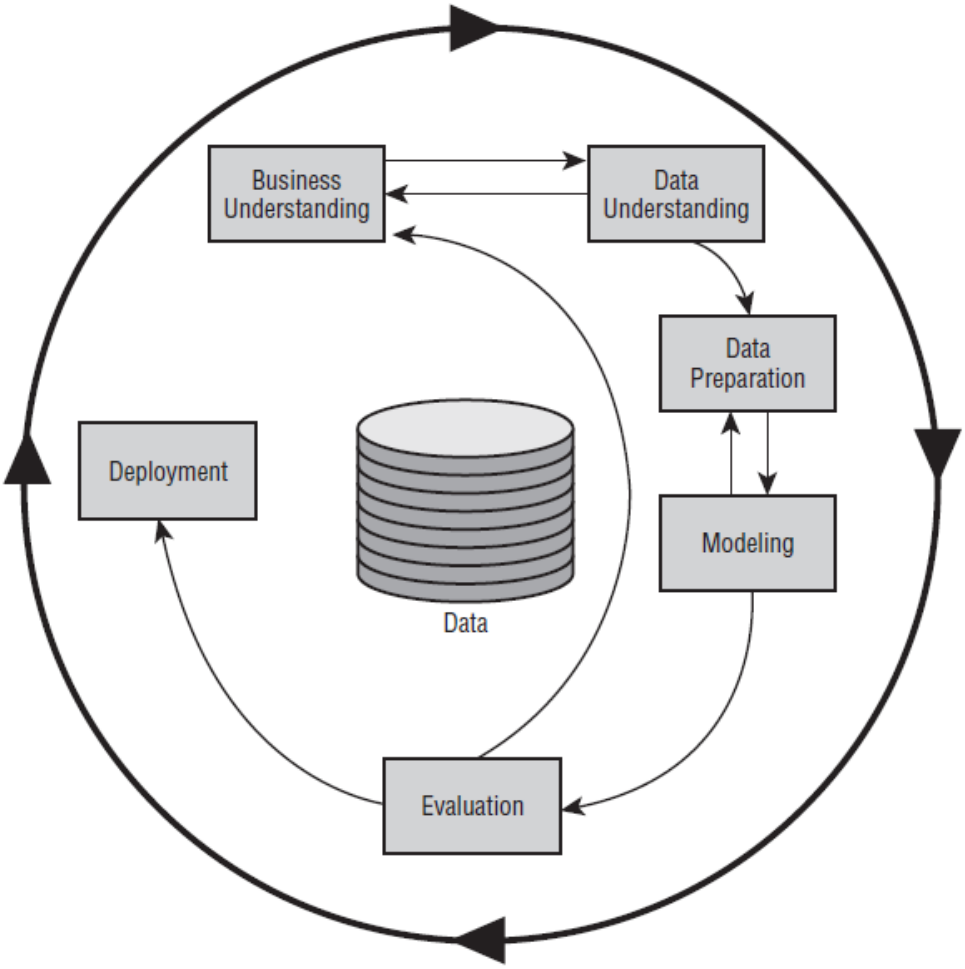

Figura 10 - Representação gráfica do CRISP-DM
(Fonte: adaptado de Provost & Fawcett)

A fase de Entendimento do Negócio foca na compreensão dos objetivos do projeto e seus requisitos numa perspectiva de negócio. Esse entendimento deve ser convertido na definição do problema de *Data Mining* e em um plano preliminar para o atingimento dos objetivos.

Na fase de Entendimento dos Dados deve haver a coleta dos dados e devem ser desenvolvidas atividades que permitam sua melhor compreensão. Análises estatísticas e técnicas de

visualização de dados são caminhos para verificar a qualidade dos dados, se de fato são relacionados com os problemas, detectar anomalias e desenvolver hipóteses.

A Preparação de Dados envolve todas as atividades necessárias para que se chegue ao *dataset* final, que é o que será utilizado na fase seguinte. Nessa fase, atividades comuns são *joins* de tabelas e cálculo de variáveis que não existem preliminarmente.

Na fase de Modelagem serão aplicados os algoritmos de *Machine Learning* e ajustados seus parâmetros para alcançar a melhor performance. Tipicamente diversos algoritmos são testados para que possam ser comparados entre si e a melhor alternativa seja selecionada. Tipicamente, diferentes algoritmos requerem diferentes formatos de dados, sendo necessário voltar a fase anterior para adaptações a cada algoritmo.

A etapa seguinte consiste em realizar a Avaliação dos modelos gerados. Eles devem ser avaliados quantitativamente a partir de métricas pré-definidas e também deve ser avaliado se o modelo de fato resolve o problema de negócio proposto. Deve ser avaliada a utilidade e aplicabilidade da solução final como um todo, ou seja, não só o modelo como sua usabilidade, rotinas necessárias para sua utilização, gestão e manutenção de ferramentas, entre outros assuntos que podem variar entre diferentes projetos. Ao final dessa etapa deve ser decidido se a solução final é adequada e será implantada.

A etapa de Implantação consiste em incluir a modelagem construída na realidade do negócio de forma a ganhar inteligência na tomada de decisão. Isso envolve criar rotinas de processamento de dados, idealmente automatizadas, rotinas de re-treino do modelo escolhido para que esteja sempre atualizado com novos dados, definição de indicadores e gestão de seu acompanhamento e mecanismos de gestão da rotina. Também é nessa etapa em que é realizada a transferência de conhecimento entre o desenvolvedor do projeto e a área de negócio cliente.

2.3.2 SEMMA

A sigla SEMMA significa *Sample, Explore, Modify, Model & Assess* e foi desenvolvido pelo *SAS Institute* para guiar o processo de *Data Mining* através de cinco etapas.

Sample é a primeira etapa do processo e consiste em selecionar os dados que serão utilizados, de forma que o resultado seja grande o suficiente para conter informações significativas, porém pequeno o suficiente para que possa ser manipulado agilmente.

Explore consiste na exploração dos dados, a procura de padrões e anomalias que auxiliem no entendimento do problema a ser resolvido.

Modify é a etapa em que as variáveis explicativas do problema serão criadas, selecionadas e transformadas, de forma que restem aquelas que melhor expliquem o problema.

Model consiste na aplicação dos algoritmos de *Machine Learning* e ajuste de seus parâmetros. Esses algoritmos construíram as relações entre as variáveis explicativas e a variável resposta.

Assess é a etapa na qual serão avaliadas as métricas de performance definidas como críticas para o projeto. Se na etapa anterior mais de uma solução tiver sido testada, nessa etapa elas serão comparadas para que seja escolhida a melhor. Também deve ser avaliada a utilidade e a confiabilidade do modelo final (AZEVEDO & SANTOS, 2008).

2.3.3 PDCA

O conceito do ciclo PDCA foi originalmente desenvolvido na década de 1930 pelo estatístico americano Walter A. Shewhart, com o propósito de atuar como um ciclo de controle estatístico de processos, podendo ser repetido continuamente sobre qualquer processo ou problema. Entretanto, foi apenas após sua aplicação pelo especialista em qualidade William Edwards Deming em seus trabalhos no Japão que o método se popularizou. (ANDRADE, 2003)

As letras que dão nome ao ciclo correspondem as palavras em língua inglesa *Plan*, *Do*, *Check*, *Act*, podendo ser traduzidas para Planejar, Executar, Verificar e Atuar. O método foi projetado para ser usado como um modelo dinâmico, de forma que seja reiniciado sempre que chegue ao fim. Ele costuma ser representado graficamente em forma circular, conforme apresentado na Figura 11, representando sua filosofia de melhoria contínua. (ANDRADE, 2003)

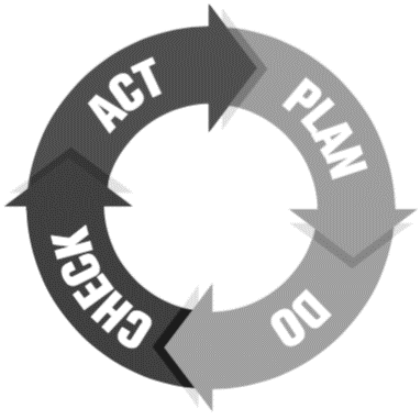

Figura 11 – Representação gráfica do ciclo PDCA
(Fonte: Creative Safety Supply)

A etapa inicial (*Plan*) pode ser considerada a mais importante do ciclo, dado que todo o restante é um desencadeamento do que ocorre nessa etapa. Tem como objetivo a criação de um plano,

traçando objetivos e elaborando os possíveis caminhos para alcançá-los. (BADIRU & AYENI, 1993)

Na etapa de execução (*Do*), os planos de ação desenvolvidos e formalizados na etapa anterior deverão ser postos em prática. O sucesso desta etapa depende da qualidade do que fora desenvolvido na etapa anterior, e deverá ser medido pelas metas elaboradas. (ANDRADE, 2003)

A terceira etapa (*Check*) consiste na verificação das ações tomadas na etapa anterior. Para que a verificação possa ser feita corretamente, é importante que as ações tomadas anteriormente sejam monitoradas e formalizadas de maneira adequada. (ANDRADE, 2003)

A quarta e última etapa (*Act*) consiste no processo de padronização daquelas ações em que se verificou eficácia na etapa anterior. O propósito disso é justamente a melhoria contínua, dessa forma esperasse que a cada iteração do ciclo os processos sejam cada vez mais eficazes. (ANDRADE, 2003)

2.3.4 DMAIC

A sigla DMAIC corresponde as palavras de língua inglesa *Define*, *Measure*, *Analyze*, *Improve* e *Control* que podem ser traduzidas para Definir, Medir, Analise, Incrementar e Controlar. A metodologia foi desenvolvida como uma evolução do ciclo PDCA, sendo dada uma maior ênfase a fase de planejamento deste. (AGUIAR, 2002)

A primeira etapa (*Define*) tem por objetivo a definição clara do escopo do projeto, definindo o problema, o cliente e suas necessidades. Deve ser definido um processo a ser melhorado e as metas a serem batidas. (CARVALHO & ROTONDARO, 2002)

Na etapa de medição (*Measure*) o desempenho do processo atual deve ser mensurado, detectando seus problemas e avaliando suas criticidades. Para tal, devem ser coletados dados de forma padronizada, permitindo que coletas similares sejam feitas no futuro para comparação. (CARVALHO & ROTONDARO, 2002)

Na etapa seguinte (*Analyze*) será feita a análise dos dados coletados anteriormente. O objetivo é identificar as causas raízes dos problemas identificados na primeira fase. Nessa etapa também será elaborado o estado futuro a ser alcançado. (CARVALHO & ROTONDARO, 2002)

A etapa de incremento (*Improve*) tem como propósito o teste e a implementação do estado futuro desenhado. Para tal, as atividades de implementação devem ser detalhadas no formato

de um plano de ação, cabendo a utilização de técnicas de gestão de projetos para maximizar as chances de sucesso. (CARVALHO & ROTONDARO, 2002)

A última etapa (*Control*) tem como objetivo a padronização das melhorias implementadas e o monitoramento das métricas de sucesso. Nessa etapa podem ser realizadas novas coletas de dados, para comparação com estado anterior e acompanhamento da efetividade das ações tomadas com o passar do tempo. (CARVALHO & ROTONDARO, 2002)

3 METODOLOGIA

Para abordagem do problema exposto nesse trabalho é importante que se utilize uma metodologia de trabalho bem definida e estruturada, de forma a garantir os melhores resultados. Segundo Terner (2008), a solução de problemas segue uma sequência lógica, tratando-se de um processo que passa pela identificação do problema, sua análise e se encerra com a tomada de decisão sobre o que deve ser feito.

A definição da metodologia para esse problema foi escolhida a partir do estudo de quatro abordagens, apresentado no Capítulo 2.3 deste trabalho. Duas dessas abordagens são mais generalistas, PDCA e DMAIC, e duas são específicas para problemas de mineração de dados, CRISP-DM e SEMMA.

É possível notar que as quatro metodologias seguem padrões similares, iniciando pela geração de hipóteses empíricas, coleta e análise de dados, propostas de melhoria e atuação sobre o problema.

Breur (2017) afirma que as metodologias propostas pelo CRISP-DM e SEMMA podem ser consideradas equivalentes ao ciclo PDCA de Deming dentro do assunto de mineração de dados. Ele afirma que essas abordagens se apoiam no método científico, no sentido de que são guiadas pela aderência aos fatos e ao teste de hipóteses, sugerindo que esses são os únicos caminhos para real transformação dos negócios.

Todas as quatro metodologias são divididas em etapas, de forma que é possível traçar paralelos entre elas. Essas correspondências são apresentadas na Tabela 2.

Tabela 2 – Correspondências entre as etapas das metodologias de trabalho

<table>
<tr><th>PDCA</th><th>DMAIC</th><th>CRISP-DM</th><th>SEMMA</th></tr>
<tr><td rowspan="5">Plan</td><td>Define</td><td>Entendimento do Negócio</td><td>-</td></tr>
<tr><td rowspan="2">Measure</td><td rowspan="2">Entendimento dos Dados</td><td>Sample</td></tr>
<tr><td>Explore</td></tr>
<tr><td rowspan="2">Analyze</td><td>Preparação dos Dados</td><td>Modify</td></tr>
<tr><td>Modelagem</td><td>Model</td></tr>
<tr><td>Do</td><td>Improve</td><td>Avaliação</td><td>Assess</td></tr>
<tr><td>Check</td><td rowspan="2">Control</td><td rowspan="2">Implantação</td><td rowspan="2">-</td></tr>
<tr><td>Act</td></tr>
</table>

(Fonte: elaborado pelo autor)

Detecção de fraudes são um típico problema de mineração de dados, sendo apontado como a mais bem estabelecida aplicação deste tipo de técnica tanto pela indústria quanto por órgãos governamentais. (PHUA, LEE, & GAYLER, 2005).

Por se tratar de um problema de mineração de dados, entende-se que é mais adequado utilizar uma metodologia de trabalho orientada a esse tipo de situação, sendo possível utilizar tanto a abordagem SEMMA quanto CRISP-DM.

A metodologia CRISP-DM foi a escolhida para guiar este projeto por trazer uma visão mais holística do processo de resolução de problemas, incluído uma etapa inicial de Entendimento do Negócio e uma etapa final de Implantação que não apresentam equivalentes na metodologia SEMMA.

Além disso, a metodologia SEMMA foi desenvolvida pela empresa SAS para guiar o processo de mineração de dados através de seu software Enterprise Miner, que não estaria disponível nesse caso, tornando-se menos interessante. (MARISCAL, MARBÁN, & COVADONGA, 2010)

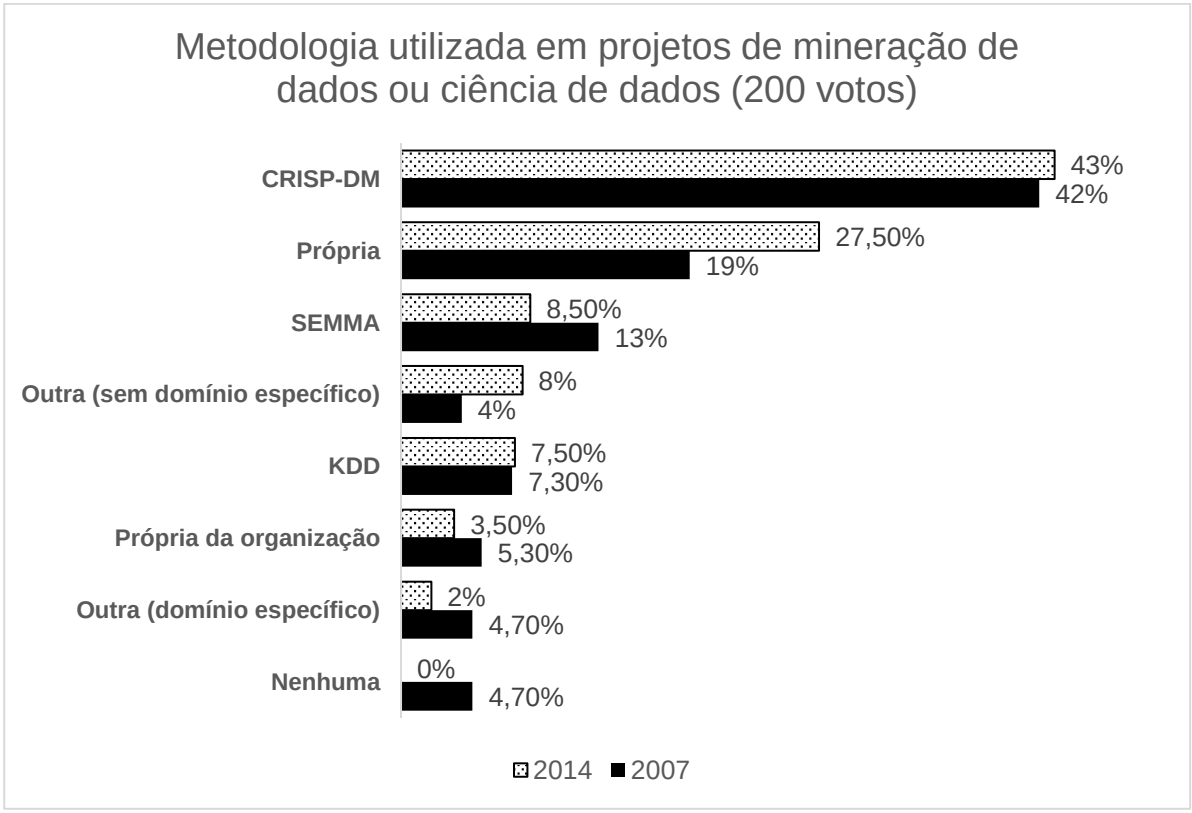

Figura 12 – Metodologias mais utilizadas em projetos de mineração de dados
(Fonte: adaptado de KDnuggets)

Outro ponto a favor do CRISP-DM é o fato de ser a metodologia mais comum empregada no desenvolvimento de projetos de mineração de dados (Figura 12), sendo utilizada em 43% dos projetos segundo estudo conduzido em 2014. Um estudo similar realizado em 2007 indicou que a abordagem era utilizada em 42%. A metodologia SEMMA foi indicada por 8,5% dos respondentes em 2014 e 13% em 2007, apresentando uma queda significativa em sua popularidade. (KDnuggets, 2014)

4 DESENVOLVIMENTO – FASE I

O projeto foi desenvolvido utilizando o *framework* de trabalho proposto pela metodologia CRISP-DM. Dada a complexidade da implantação do mesmo, o projeto foi dividido em duas fases, sendo a Fase I composta pelas etapas que vão do Entendimento do Negócio a Avaliação. A Fase II tratou exclusivamente da etapa de Implantação.

4.1 Entendimento do Negócio

Ao receber um cheque, um indivíduo pode optar por trocá-lo diretamente por dinheiro, o que deve ser feito obrigatoriamente em uma agência da instituição financeira do emissor do cheque, seguindo eventuais restrições com relação ao valor do cheque, à necessidade de recorrer a uma agência específica e à necessidade de reserva de numerário.

A outra opção é realizar o depósito do valor do cheque em uma conta corrente de sua escolha, ou seja, um cheque pode ser depositado em uma conta corrente de um banco diferente do emissor. Feito o depósito, uma série de etapas que constituem um fluxo físico e outro de informação se seguem, no chamado processo de compensação, como apresentado na Figura 13. É importante notar que as instituições financeiras A e B representadas podem, ou não, serem a mesma. O processo de conferência de cheques sempre será realizado pela mesma instituição do emissor do cheque, enquanto o processo de compensação física do cheque será realizado pela instituição do depositante do cheque.

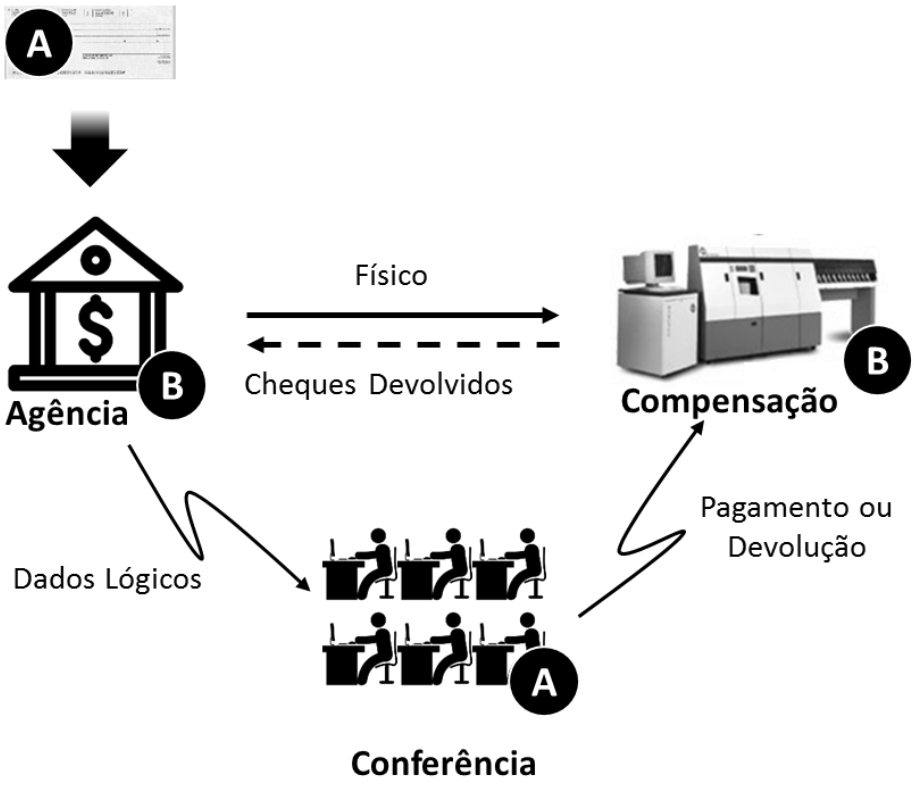

Figura 13 – Representação do fluxo físico e de informação do processo de compensação de cheques

(Fonte: elaborado pelo autor)

Dado que a agência de um determinado banco pode receber depósitos de cheques de qualquer outra instituição, é necessário um responsável pela consolidação e redistribuição das informações.

Ao final do dia, cada agência captura a imagem e cadastra os dados lógicos de todos os cheques lá depositados durante o dia. Esses dados são enviados ao Banco do Brasil, que consolida a recepção dos dados de todas as agências de todos os bancos brasileiros.

O Banco do Brasil processa os dados e redistribui, enviando para cada instituição os cheques emitidos por seus clientes, independentemente de onde tenham sido depositados. Esse envio ocorre na madrugada, por volta das duas horas da manhã do dia seguinte ao depósito. Esse fluxo de informação é detalhado na Figura 14 e a sequência e momento dos eventos é apresentado na Figura 15.

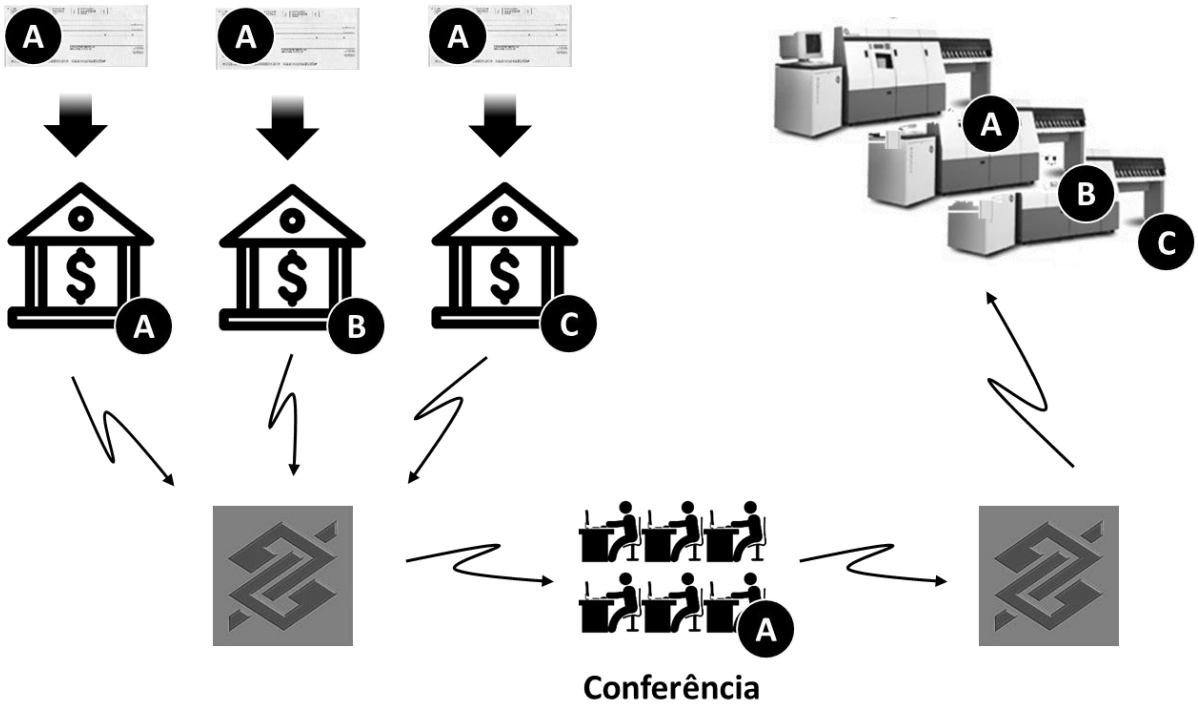

Figura 14 – Detalhamento do fluxo de informação e da interação entre as instituições financeiras no processo de compensação de cheques

(Fonte: elaborado pelo autor)

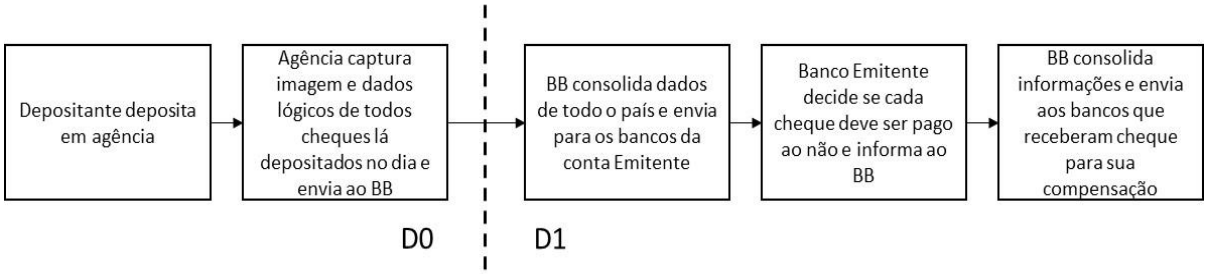

Figura 15 – Fluxo de informação no processo de compensação de cheques

(Fonte: elaborado pelo autor)

A conferência de cheques consiste na análise da imagem da folha do cheque para que sejam avaliados o correto preenchimento do cheque, a autenticidade da assinatura e a autenticidade da folha em si. O não cumprimento de algum desses fatores acarreta na devolução do cheque. Um cheque também pode ser devolvido função de outros motivos, como, por exemplo, a inexistência de fundos para pagamento. Entretanto, nesse caso, a devolução ocorre de forma sistêmica, independente do resultado da conferência.

Dada a natureza manual da operação e o alto volume diário de cheques, a operação de conferência de cheques torna-se muito custosa para instituição financeira, com mais de 100 pessoas alocadas apenas na operação. Além disso, quando a operação falha e paga um cheque

que deveria ser devolvido, o banco é obrigado a ressarcir o cliente acarretando em perdas financeiras. O outro tipo de erro, devolver um cheque autêntico, desgasta a imagem do banco com seus clientes.

Uma possível opção para a redução dos custos operacionais é conferir apenas os cheques que oferecem um maior risco de perda para o banco. Esse risco é função do valor do cheque e da probabilidade de irregularidade, que será aqui tratada como fraude.

Após entrevistas com colaboradores da instituição financeira, foram mapeadas as situações que definem se um cheque deve considerado uma fraude:

- Cheques pagos pela operação, que foram contestados pelo cliente e houve ressarcimento,
- Cheques devolvidos pela operação por motivos que constituem fraude: Folha clonada e Assinatura copiada.

Dessa forma, os objetivos propostos para o projeto foram o desenvolvimento de um modelo capaz de avaliar a probabilidade de fraude de cada cheque e com seu valor calcular o risco de perda, criar uma lógica de priorização da conferência dos cheques e determinar o volume de conferência da operação, atingindo o nível ótimo entre custo operacional e perdas com fraudes.

É importante que tal modelo seja capaz de ser aplicado aos cheques que chegam diariamente em pouco tempo, pois há uma curta janela de tempo entre o recebimento do arquivo com os cheques do dia enviado pelo Banco do Brasil, às duas horas da manhã, e o início da operação de conferência de cheques, às seis da manhã.

Os recursos disponíveis para a realização do projeto são os softwares SAS, R e Excel. O SAS poderá ser utilizado para acesso a bases de dados e manipulação de dados e foi escolhido por ser a plataforma padrão adotada pela empresa cliente do projeto. O software de R será utilizado para a modelagem de *Data Mining* e foi escolhido por ser um software *open source* voltado para esse tipo de aplicação, sendo um dos mais utilizados na indústria. O Excel foi escolhido como interface do usuário por ser de uso diário dos clientes do projeto.

Através de diversas entrevistas e *brainstormings* com colaboradores da área cliente e outras áreas correlatas foi feito um levantamento dos possíveis fatores que possam explicar a fraude em cheques, esses fatores foram agrupados em categorias a partir de sua classificação conforme expectativas de sua importância para o modelo e dificuldade de obtenção de dados para seu cálculo.

Esse levantamento é apresentado na Tabela 3. O objetivo de tal levantamento é direcionar esforços para a obtenção de dados, focando naqueles que se espera uma maior correlação com a fraude e que podem ser obtidos com maior rapidez. Esse levantamento é feito com objetivo exploratório e exaustivo, de forma que não necessariamente todos os dados serão levantados e/ou úteis.

Tabela 3 – Levantamento de possíveis variáveis explicativas para a fraude em cheques

Variável	Dificuldade de Obtenção	Importância Estimada
Emitente já fraudado	Média	Alta
Depositante já recebeu fraude	Média	Alta
Comportamento de emissão diária	Alta	Alta
Status da conta	Baixa	Alta
Valor do cheque	Baixa	Média
Histórico de insuficiência de saldo	Média	Média
Conta Digital	Média	Média
Comportamento de depósito diária	Alta	Alta
Renda do emitente	Média	Média
Gênero do emitente	Baixa	Média
Tipo Pessoa do emitente (PF/PJ)	Baixa	Média
Idade da conta do emitente	Baixa	Média
Cheque reapresentado	Baixa	Média
Idade do emitente	Baixa	Média
Estado civil do emitente	Baixa	Média
Escolaridade do emitente	Baixa	Média
Conta conjunta do emitente	Baixa	Média
Histórico de devoluções do emitente	Baixa	Média
Histórico de devoluções do depositante	Baixa	Média
Perfil de uso (quantidade)	Baixa	Alta
Meio de entrega	Média	Alta
Segmento comercial do emitente	Baixa	Alta
UF - emitente	Baixa	Baixa

UF – depósito	Baixa	Baixa
Cidade - emitente	Média	Baixa
Cidade – depósito	Média	Baixa
CEP - emitente	Alta	Média
CEP - depósito	Alta	Média
Data de fabricação do cheque	Média	Média
Variação na numeração dos cheques	Alta	Alta
Relacionamento emitente/depositante	Média	Alta
Banco de depósito	Baixa	Média
Emitente funcionário do banco	Média	Baixa
Cadastro biométrico	Alta	Média
Histórico de valor do emitente	Média	Alta
Histórico de valor do depositante	Média	Alta
Empresa transportadora	Alta	Média
Idade do depositante	Alta	Média
Estado civil do depositante	Alta	Média
Escolaridade do depositante	Alta	Média
Conta conjunta do depositante	Alta	Média
Histórico de devoluções do depositante	Alta	Média

4.2 Entendimento dos Dados

A base de dados que será utilizada como ponto de partida para o desenvolvimento do projeto contém todos os cheques que passaram pela operação de conferência, sendo cada linha um cheque diferente. Por limitações de infraestrutura da área, esses dados são armazenados em arquivos quinzenais do MS Access, sendo necessária sua consolidação em uma tabela única que, pelo alto volume dados, só seria suportada por um servidor no SAS.

Nessa fase do projeto também foram levantadas outras fontes de dados, que forneçam as informações adicionais que serão utilizadas no cálculo das variáveis explicativas, dessa forma um assunto a ser tratado na etapa seguinte do desenvolvimento é integrar as múltiplas fontes de

dados em uma única tabela, na qual cada linha representará um cheque e suas informações, tanto explicativas quanto a variável resposta, ou seja, se é ou não uma fraude.

Para um melhor entendimento das variáveis e eventuais correlações entre elas e o evento de interesse, foram utilizadas análises bivariadas, utilizando técnicas de visualização de dados para evidenciar tais correlações. As análises consistem em comparar cada variável explicativa com a resposta, e avaliar o volume de dados e o índice de fraude de cada nível da variável explicativa. Entende-se por níveis da variável explicativa os possíveis valores que ela pode assumir.

A Figura 16 é um exemplo da análise bivariada em função de uma variável explicativa e da variável resposta. É possível observar como o índice de fraude é destoante entre seus níveis, de forma que é uma variável interessante para ser incorporada ao modelo.

Foi desenvolvido um material com análises desse tipo para todas as variáveis estudas nessa etapa do projeto, sendo uma entrega para o cliente que permite um melhor entendimento dos fatores que impactam a probabilidade de fraude. O desenvolvimento desse material permitiu que a operação tomasse ações de *quick-win*, gerando melhorias e captura de ganhos, como redução de perdas com fraude devido a melhor seleção de cheques para conferência, enquanto o projeto ainda estava em desenvolvimento.

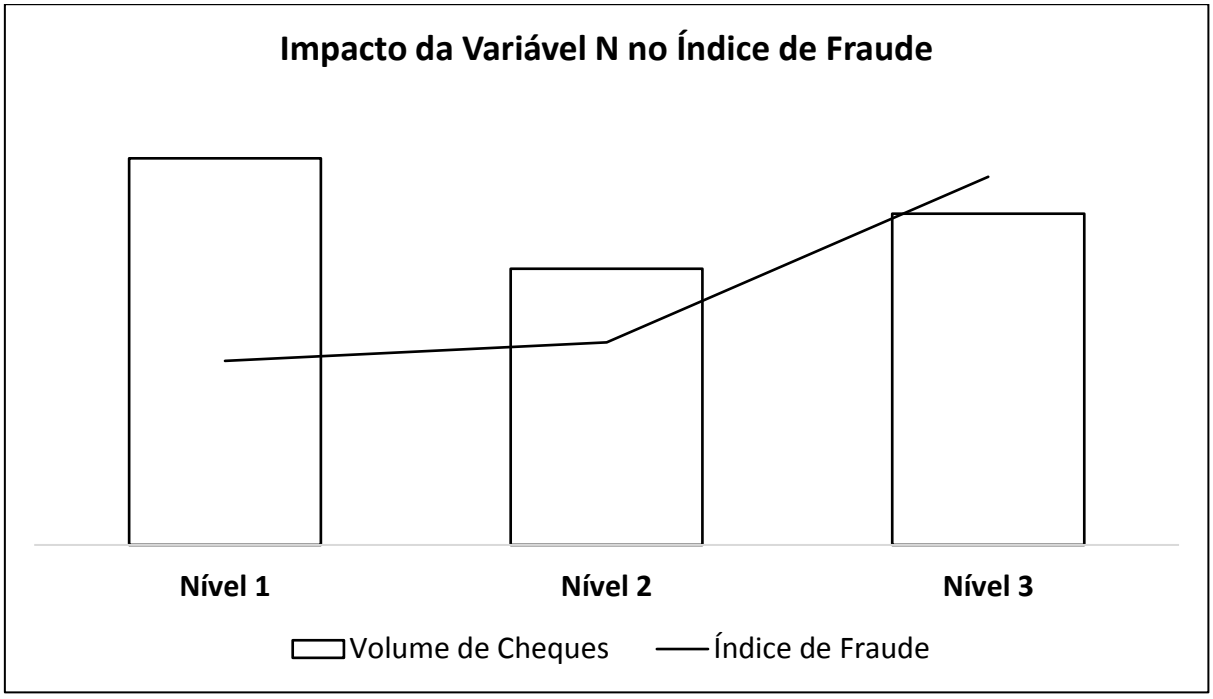

Figura 16 – Exemplo de análise bivariada, analisando a variável resposta em função de uma variável explicativa

(Fonte: elaborado pelo autor)

Por questões de confidencialidade não serão apresentados os resultados das análises bivariadas. Também não será explicitada a variável em questão na Figura 16.

4.3 Preparação dos Dados

A partir das análises construídas na fase anterior, foi possível avaliar o poder explicativo das variáveis bem como eventuais restrições técnicas com relação a seu acesso e disponibilidade, tendo como referencial a necessidade de dados que possam ser acessados de forma rápida com o modelo em operação.

A partir dos dados levantados, foi identificada a possibilidade de se construir novas variáveis derivadas das originais, como, por exemplo, a existência ou não de relacionamento comercial entre as contas emitente e depositante e variações comportamentais dos mesmos, seja com relação a valor dos cheques ou a quantidade diária de cheques emitidos e depositados. Após sua construção e análises semelhantes às construídas para as outras variáveis foram classificadas da mesma maneira para avaliar sua inclusão no modelo final.

Todas as tabelas foram integradas, gerando a tabela final que servirá de insumo para os algoritmos que serão testados na fase seguinte, como exemplificado na Tabela 4. Ela contém todos os cheques do período selecionado para o estudo, a variável resposta que indica se o cheque foi ou não classificado como uma fraude e as variáveis explicativas selecionadas nas fases anteriores.

Tabela 4 - Exemplo genérico de tabela de dados final

ID Cheque	**Variável Resposta**	**Variável Explicativa 1**	**...**	**Variável Explicativa N**
1	Sim	Var1_Nível1	...	VarN_Nível1
...	...	...	...	...
n	Não	Var1_NívelN	...	VarN_NívelN

(Fonte: elaborado pelo autor)

4.4 Modelagem

Nessa fase ocorre a construção do modelo que será aplicado aos cheques futuros. A construção do modelo se dá através da aplicação de um algoritmo a um conjunto de dados. O modelo, uma

vez construído, pode ser aplicado a novos conjuntos de dados, prevendo o comportamento da variável resposta para os elementos deste novo conjunto. Essa dinâmica é apresentada na Figura 17.

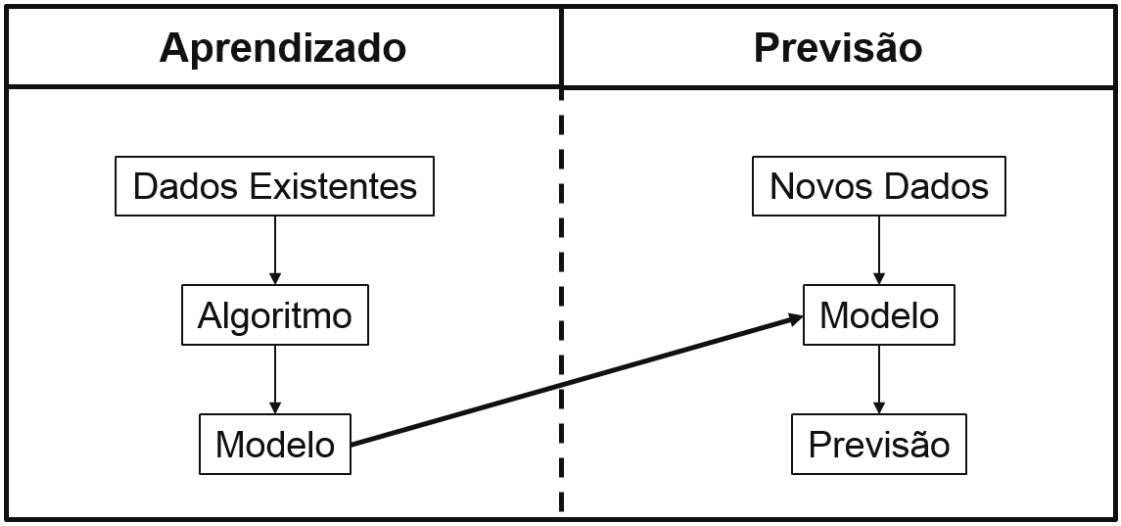

Figura 17 - Dinâmica de aprendizagem e aplicação do modelo

(Fonte: elaborado pelo autor)

A existência de diferentes algoritmos aplicáveis ao problema em questão bem como diferentes possibilidades de arranjos dos dados, levam a um cenário em que diversas estratégias podem ser testadas e avaliadas para que se selecione a mais apropriada ao modelo. As variações mencionadas se referem-se:

- Ao algoritmo escolhido;
- A necessidade de balanceamento dos dados;
- Ao volume de dados necessário para o treinamento, também chamado de aprendizado.

Com relação ao algoritmo, foram escolhidas três técnicas para teste:

- Árvore de Decisão;
- *Random Forest*.

Com relação ao balanceamento decidiu-se por realizar o teste de três formas diferentes: sem balanceamento, utilizando técnicas de *downsampling* para alcançar um índice de fraude de 30% e utilizando o mesmo tipo de técnica para alcançar um índice de fraude de 50%.

Também se decidiu avaliar o impacto do volume de dados utilizado no treinamento. Grandes volumes de dados levam a um grande esforço computacional e consequentemente largos períodos de processamento, é desejado avaliar o *trade-off* entre o tempo gasto no treinamento

e sua performance. Dessa forma, propõe-se realizar os treinamentos e testes com quatro quantidades de linhas diferentes: 5 mil, 25 mil, 50 mil e 100 mil linhas.

Todos os arranjos planejados inicialmente são apresentados na Tabela 5.

Tabela 5 - Arranjos planejados para escolha da modelagem

Nome do Arranjo	Algoritmo	Balanceamento	Volume de Dados
AD_N_5k	Árvore de Decisão	N	5 mil
AD_N_25k	Árvore de Decisão	N	25 mil
AD_N_50k	Árvore de Decisão	N	50 mil
AD_N_100k	Árvore de Decisão	N	100 mil
AD_30_5k	Árvore de Decisão	30%	5 mil
AD_30_25k	Árvore de Decisão	30%	25 mil
AD_30_50k	Árvore de Decisão	30%	50 mil
AD_30_100k	Árvore de Decisão	30%	100 mil
AD_50_5k	Árvore de Decisão	50%	5 mil
AD_50_25k	Árvore de Decisão	50%	25 mil
AD_50_50k	Árvore de Decisão	50%	50 mil
AD_50_100k	Árvore de Decisão	50%	100 mil
RF_N_5k	*Random Forest*	N	5 mil
RF_N_25k	*Random Forest*	N	25 mil
RF_N_50k	*Random Forest*	N	50 mil
RF_N_100k	*Random Forest*	N	100 mil
RF_30_5k	*Random Forest*	30%	5 mil
RF_30_25k	*Random Forest*	30%	25 mil
RF_30_50k	*Random Forest*	30%	50 mil
RF_30_100k	*Random Forest*	30%	100 mil
RF_50_5k	*Random Forest*	50%	5 mil
RF_50_25k	*Random Forest*	50%	25 mil
RF_50_50k	*Random Forest*	50%	50 mil
RF_50_100k	*Random Forest*	50%	100 mil

Os testes dos modelos consistiram em selecionar amostras aleatórias para a realização do treinamento do modelo e para seu teste. Isso permite avaliar com maior precisão qual será o resultado do modelo em uma operação real, pois a aplicação ocorrerá em dados nunca vistos pelo algoritmo de treinamento.

A base de testes consistirá em uma tabela de dados com um milhão de transações. Tal volume permite minimizar a aleatoriedade dos resultados, e é viável, pois a aplicações de modelos desse tipo são muito rápidas, inferiores a um minuto. Será utilizada a mesma base de testes para todos os arranjos experimentais propostos, tornando a comparação entre seus resultados adequada.

Foi desenvolvido um código na linguagem R para ler as bases de dados, treinar o modelo preditivo baseado em um parâmetro que define o algoritmo utilizado, aplica o modelo resultante na base de testes e calcula as métricas de avaliação que interessam ao problema e serão detalhadas no próximo subcapítulo deste trabalho. Por questões de confidencialidade não será apresentado o código escrito pelo autor, apenas sua estrutura em pseudocódigo na Figura 18. Também será apresentada a estrutura da rotina de geração das bases de treino e teste.

Figura 18 - Estrutura do Código de treino e teste do modelo

(Fonte: elaborado pelo autor)

4.5 Avaliação

Problemas de detecção de fraude são classicamente muito desbalanceados, isto é, uma das classes, no caso a de transações legítimas, possui muito mais representantes do que a outra. No caso específico estudado essa proporção é superior a 99%.

Nesse tipo de situação, métricas de desempenho como a acurácia podem ser enganosas, já que um modelo que aponte todas as transações como pertencentes a classe predominante teria uma acurácia superior a 99%, considerada altíssima. Métricas como a Sensibilidade e a Especificidade podem contornar esse problema, mas somente se olhadas de forma conjunta. Não obstante, é importante notar que os algoritmos de *Machine Learning* calculam *scores* entre e 0 e 1, que indicam a probabilidade de pertencimento a determinada classe. Essas métricas dependem da definição de um limite pré-estabelecido que determine se cada cheque será considerado uma fraude ou não. Por padrão esse limite é 0.5, mas no projeto em questão

estamos mais interessados nos *scores*, do que em separar os cheques entre os que acreditamos que serão fraudes ou não. Isso porque o modelo será utilizado para priorizar a conferência dos cheques em uma operação com capacidade fixa. Isto é, serão conferidos os cheques de maior risco.

Uma alternativa para visualizar a qualidade do modelo como um todo são as curvas ROC (*Receiver Operating Characteristic*), que mostram como a Sensibilidade e Especificidade variam com relação a essa escolha. Nota-se que cada ponto da curva representa uma matriz de confusão diferente (ver Figura 9) (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

Para consolidar a qualidade do modelo em um único número será utilizada a área sob a curva ROC, medida chamada de AUC. Esse indicador pode variar entre 0 e 1, sendo um indicador do tipo quanto maior melhor. Um modelo de classificação aleatório possui AUC de 0,5. Dessa forma, modelos com AUC inferior a esse valor não acrescentam nenhuma inteligência ao problema. (BHATTACHARYYA, JHA, THARAKUNNEL, & WESTLAND, 2011).

Outro fator de interesse é o tempo de processamento necessário para o treinamento do modelo. Alguns modelos podem apresentar alto desempenho explicativo, porém dada sua complexidade exigem um esforço computacional muito grande, sendo desejável avaliar esse *trade-off*.

A Tabela 6 apresenta a AUC de cada modelo e tempo, em segundos, necessário para seu treinamento.

Tabela 6 – Métricas de desempenho dos arranjos propostos

Nome do Arranjo	AUC	Tempo de Treinamento
AD_N_5k	0,5000	9
AD_N_25k	0,4088	10
AD_N_50k	0,3837	10
AD_N_100k	0,4457	10
AD_30_5k	0,7060	12
AD_30_25k	0,7060	31
AD_30_50k	0,7060	52
AD_30_100k	0,7060	123
AD_50_5k	0,7589	13
AD_50_25k	0,7963	37

AD_50_50k	0,7963	58
AD_50_100k	0,7963	119
RF_N_5k	0,7451	63
RF_N_25k	0,7879	284
RF_N_50k	0,7942	748
RF_N_100k	0,7967	2281
RF_30_5k	0,8621	110
RF_30_25k	0,8746	674
RF_30_50k	0,8763	1459
RF_30_100k	0,8792	3187
RF_50_5k	0,8675	72
RF_50_25k	0,8765	508
RF_50_50k	0,8806	1343
RF_50_100k	0,8836	2908

4.5.1 Impacto do Algoritmo

Analisando os dados é possível observar que a performance dos modelos construídos utilizando *Random Forest* foi bem superior aos que utilizaram Árvores de Decisão. Entretanto essa vantagem vem associada a treinamento mais longos, como pode ser observado na Figura 19.

Mesmo nos cenários mais complexos, o tempo de treinamento não foi um problema. O arranjo com maior tempo de treinamento foi utilizando o algoritmo *Random Forest*, com balanceamento de 30% e 100 mil linhas, com um tempo de 3187 segundos, ou pouco mais de 53 minutos. Esse tempo, ainda que longo em comparação com outros modelos, não representa uma dificuldade à operacionalização do modelo e suas rotinas de re-treino. Portanto, faz sentido utilizar como critério de decisão apenas a performance do modelo.

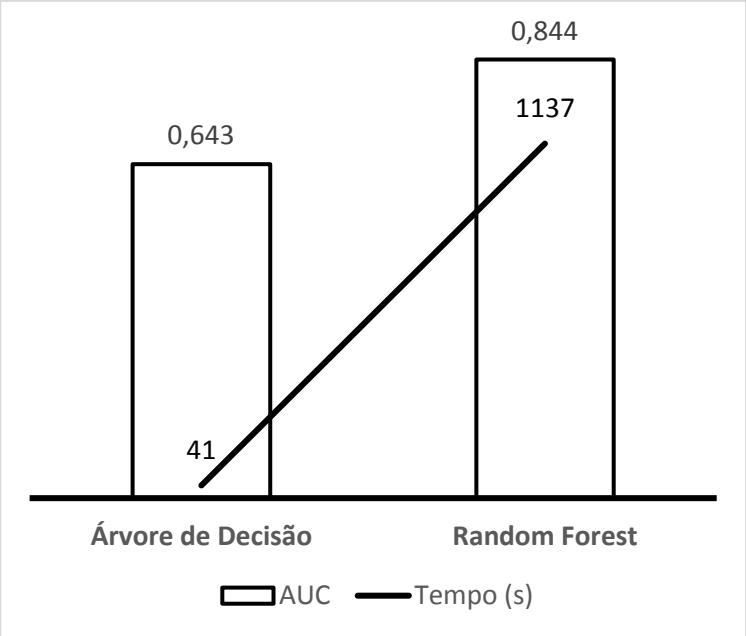

Figura 19 – Desempenho médio dos algoritmos

(Fonte: elaborado pelo autor)

Na Figura 20 é possível notar que o balanceamento da amostra traz ganhos de performance relevantes para o modelo. Isso acontece, pois, o índice de fraude é inferior a 1%, tornando a tarefa de encontrar características que distingam as fraudes dos demais cheques, extremamente difícil. Isso ocorre porque haverá poucos exemplos cheques fraudulentos na base de dados em comparação com os cheques legítimos.

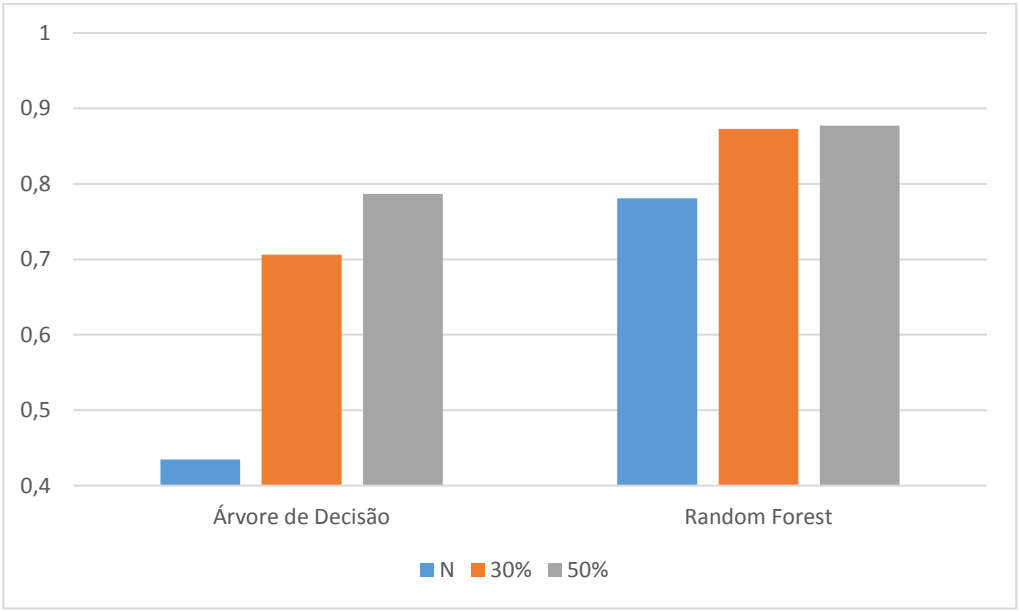

Figura 20 – Impacto do balanceamento no desempenho dos modelos

(Fonte: elaborado pelo autor)

Na Figura 21 observa-se que o volume de dados na base de treinamento não impactou diretamente a performance dos modelos que utilizaram o algoritmo de Árvore de Decisão. Já no caso do *Random Forest* houve um impacto, especialmente para tamanhos menores de base.

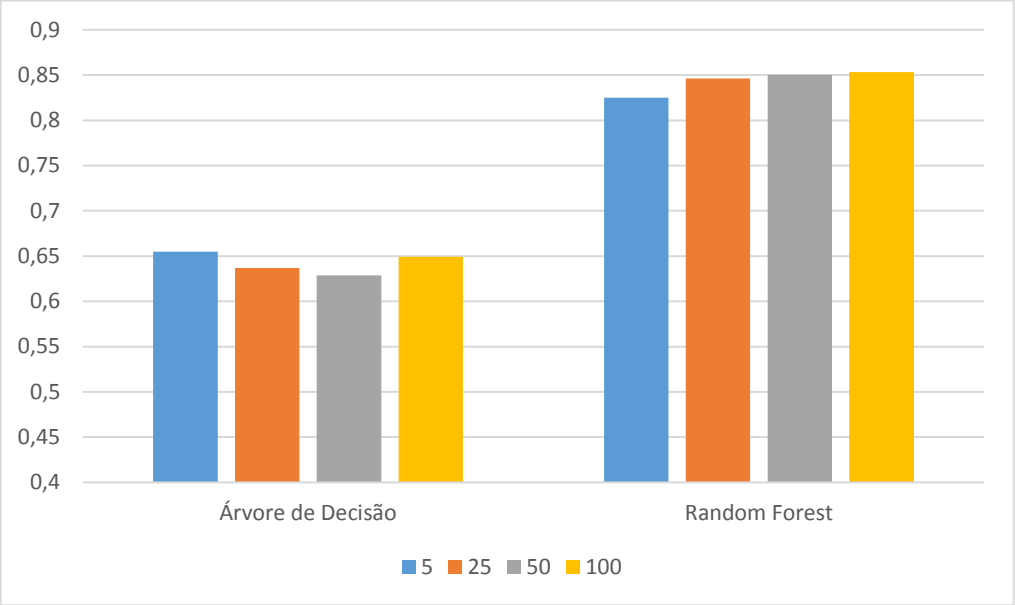

Figura 21 – Impacto do volume de dados da base de treinamento na performance (Fonte: elaborado pelo autor)

Ao analisarmos a performance em formato de mapa de calor (Figura 22), notamos que de forma geral a performance do *Random Forest* é muito superior à da Árvore de Decisão.

Com relação ao balanceamento, fica clara a necessidade da utilização da técnica de downsampling em problemas altamente desbalanceados, como é o caso da detecção de fraudes. Para ambos os algoritmos há um salto de performance ao realizar o balanceamento de 30% e também há ganhos ao utilizar um balanceamento de 50%.

Com relação ao volume de dados utilizados no treinamento, nota-se que o impacto não é claro quando se trata da Árvore de Decisão. Para cada um dos três níveis de balanceamento a performance piorou, foi indiferente, e melhorou com o aumento do volume de dados. Já no caso do *Random Forest*, um aumento no volume de dados sempre trouxe ganho em performance.

Árvore de Decisão			
	N	30%	50%
5	0,500	0,706	0,759
25	0,409	0,706	0,796
50	0,384	0,706	0,796
100	0,446	0,706	0,796

Random Forest			
	N	30%	50%
5	0,745	0,862	0,868
25	0,788	0,875	0,877
50	0,794	0,876	0,881
100	0,797	0,879	0,884

Figura 22 – Mapa de calor da performance de todos os arranjos

(Fonte: elaborado pelo autor)

Considerando os resultados obtidos, conclui-se que o melhor algoritmo para este problema é o *Random Forest*. A Figura 23 permite visualizar graficamente os efeitos citados acima no caso desse algoritmo. Além disso, ficou claro que se deve optar por realizar o balanceamento através da técnica de *downsampling*, sendo escolhido o parâmetro de 50% de balanceamento pois apresentou resultados ligeiramente superiores. Além disso, dado que o tempo de treinamento não se mostrou um fator de risco a operacionalização do modelo, foi escolhido utilizar uma base de dados de treinamento de 100 mil linhas.

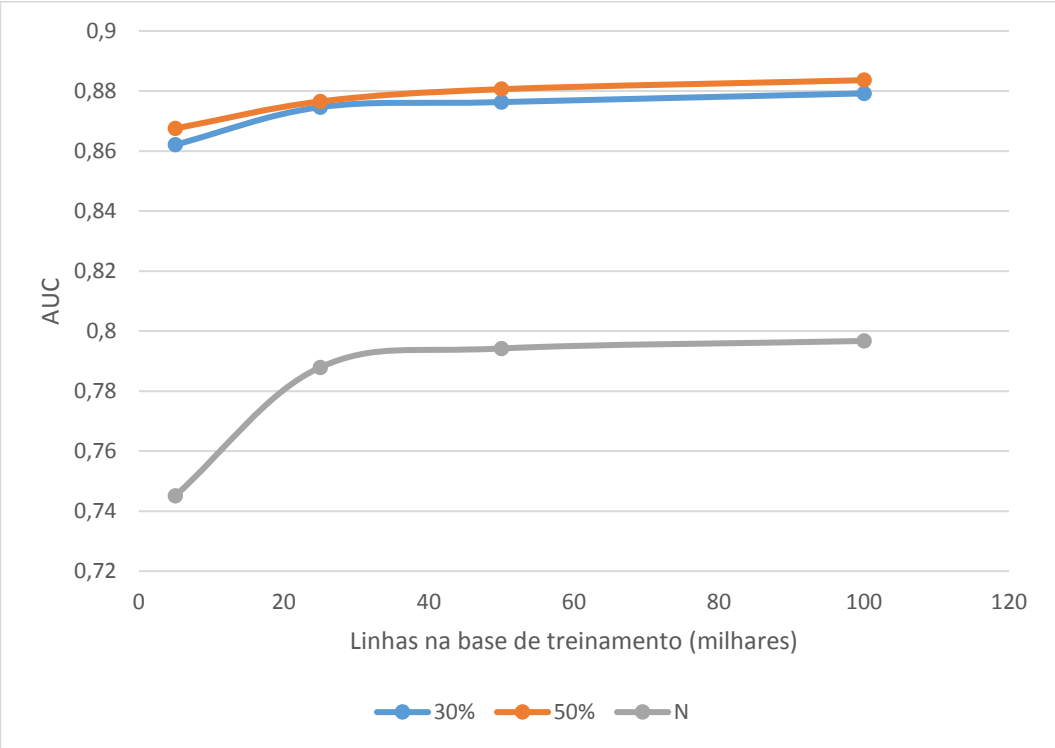

Figura 23 – Efeito do volume de dados e do balanceamento na performance do algoritmo *Random Forest*

(Fonte: elaborado pelo autor)

5 DESENVOLVIMENTO – FASE II

Seguindo a estrutura proposta pela metodologia CRISP-DM, a segunda fase do projeto consistiu em sua implantação na área de conferência de cheques. O principal desafio enfrentado na fase de implantação se deu pela curta janela de tempo entre a recepção dos dados e o início da operação de conferência de cheques.

As informações e as imagens dos cheques são enviadas pelo Banco do Brasil entre as duas e três horas da manhã, e o processo de conferência de cheques ocorre as seis horas da manhã. De forma que, a depender do dia, há apenas três horas para que seja tomada a decisão sobre quais cheques deverão ser conferidos naquele dia.

Para que tal decisão seja tomada com base no modelo criado na Fase I é necessário que nesse intervalo de tempo sejam realizadas as seguintes atividades:

1. Calcular todas as variáveis explicativas para cada um dos cheques
2. Aplicar o modelo, calculando o *score* de cada um deles
3. Definir os cheques que serão conferidos
4. Carregar a lista de cheques que irão para a conferência no sistema

Dentre as atividades supracitadas, a de cálculo de variáveis explicativas para todos os cheques do dia é a que apresenta a maior dificuldade de operacionalização. Isso ocorre porque dado o grande volume de dados, o cálculo de variáveis pode se tornar excessivamente longo e não ser factível para a operação.

O cálculo de variáveis envolve o cruzamento da base de dados que contém os cheques do dia com a base de dados cadastrais dos clientes do banco, com o histórico de todos os cheques depositados nos últimos doze meses, entre outras, e envolve diversos cálculos e diferentes níveis de agregação. A Tabela 7 indica a ordem de grandeza dessas tabelas. Por questões de confidencialidade os números exatos não serão disponibilizados.

Tabela 7 – Bases de dados envolvidas no cálculo e seus volumes

Base de dados	Número de linhas
Cheques do dia	Centenas de milhares
Histórico de cheques	Centenas de milhões
Dados Cadastrais	Dezenas de milhões
Dados Cadastrais - 2	Dezenas de milhões
Produção de Talões	Milhões

Para que fosse possível executar a atividade de cálculo de variáveis em tempo hábil foram criadas rotinas de cálculo prévio de variáveis, uma mensal e outra diária. Dessa forma, quando o cálculo de variáveis for ocorrer na madrugada que antecede a conferência, o esforço computacional, e consequentemente o tempo gasto, para a conclusão da atividade será bem menor, variando entre sessenta e noventa minutos.

Além da criação das rotinas e do desenvolvimento de códigos e programas nelas utilizados, a fase de implantação também se caracterizou pela definição dos papéis e responsabilidades dos analistas da área e treinamentos para esse mesmo público.

Os analistas serão os responsáveis pela execução das rotinas que serão apresentadas nas seções abaixo, sendo a maior parte delas automáticas de forma que requerem pouco tempo dedicado do analista, que basicamente inicia o processo e se certifica que ele ocorreu da forma correta. Em caso de falhas haverá uma área de sustentação de projetos responsável por corrigir eventuais problemas, sendo que aqueles que afetem a rotina da operação serão tratados em caráter de urgência.

Com relação aos treinamentos, o principal objetivo era contextualizar os analistas da área sobre os métodos e mecanismos que serão utilizados no novo modelo. Uma característica negativa dos algoritmos de *Machine Learning* é a de poderem ser interpretados como caixas-pretas, ou seja, é difícil para uma pessoa racionalizar suas decisões. Dado isso, o entendimento dos princípios e conceitos envolvidos nessas técnicas é essencial para que a área cliente confie no modelo e se sinta confortável em seguir seus resultados na operação.

5.1 Rotina de cálculo mensal

A rotina de cálculo mensal tem como objetivo calcular uma série de variáveis para que estas possam ser acessadas de forma mais ágil no dia-a-dia da operação. A rotina consiste de três etapas:

1. Execução de um programa em um servidor SAS
2. Transferência de dados para o servidor da área
3. Execução de um programa em um segundo servidor SAS

A Etapa 2 é necessária pois a estrutura de dados da empresa é dividida em mais de um servidor, e há dados utilizados pelo projeto em um servidor diferente daquele utilizado pela área cliente, no caso a área de conferência de cheques. Não é possível realizar cruzamento de dados que estejam em servidores diferentes, sendo necessária a transferência através de um software específico.

As etapas são automáticas, porém a ativação de cada uma delas requer interação humana. A execução da atividade ficou sob a responsabilidade de um analista da área, que além de ativar as rotinas, também as supervisiona, informando a área de sustentação de projetos em caso de falhas. O tempo total da rotina mensal, incluindo as três etapas, é de aproximadamente seis horas. Durante esse período, é possível que o analista realize suas outras atividades, dado que o programa não onera sua estação de trabalho.

5.2 Rotina de cálculo diária

A rotina de cálculo diária possui o mesmo objetivo da rotina mensal, entretanto é dedicada a variáveis cuja tempestividade exigem uma frequência de atualização maior. Essa rotina consiste de duas etapas:

1. *Upload* cinco bases de dados no servidor SAS
2. Execução de um programa no servidor SAS

A Etapa 1 não é automática, exigindo a alocação de um analista para a extração dessas cinco bases de seus sistemas e seu *upload* no servidor. A extração das bases leva aproximadamente dez minutos e seus *uploads* totalizam um período de sessenta minutos.

A Etapa 2 é automática, entretanto exige ativação manual de um analista. É necessária supervisão, sendo que problemas deverão ser reportados a área de sustentação de projetos. O programa leva cinco horas para sua execução completa, nesse período é possível que o analista realize suas outras atividades.

5.3 Rotina de aplicação do modelo

A rotina de aplicação do modelo ocorre diariamente e inicia assim que o arquivo com os dados lógicos dos cheques depositados no dia anterior é enviado pelo Banco do Brasil. A rotina consiste de quatro etapas, sendo elas:

1. Extração da base de dados com os cheques depositados no dia anterior
2. Execução de um programa no servidor SAS
3. Execução do modelo de *Machine Learning*
4. *Upload* da lista de cheques a serem conferidos na operação

A Etapa 1 é manual e leva aproximadamente quinze minutos. Seu objetivo é gerar um arquivo de texto contendo os cheques depositados no dia anterior que será acessado pelo programa em SAS na etapa seguinte.

O programa da Etapa 2 acessa o arquivo de texto e cruza com as bases de variáveis pré-calculadas, gerando uma tabela em que cada linha representa um cheque e as colunas são o ID de cada cheque e todas as variáveis contidas no modelo desenvolvido na Fase I. Essa tabela é salva como arquivo de texto em um local pré-determinado na rede. Essa etapa dura aproximadamente noventa minutos.

Na Etapa 3 é ativado o *script* em R que realiza a aplicação do modelo de *Machine Learning* desenvolvido na Fase I deste projeto, e tem sua execução completa em aproximadamente dez minutos. A ativação é feita através de interface desenvolvida em Excel e VBA. Por questões confidencialidade o script não será disponibilizado, apenas sua estrutura apresentada na Figura 24.

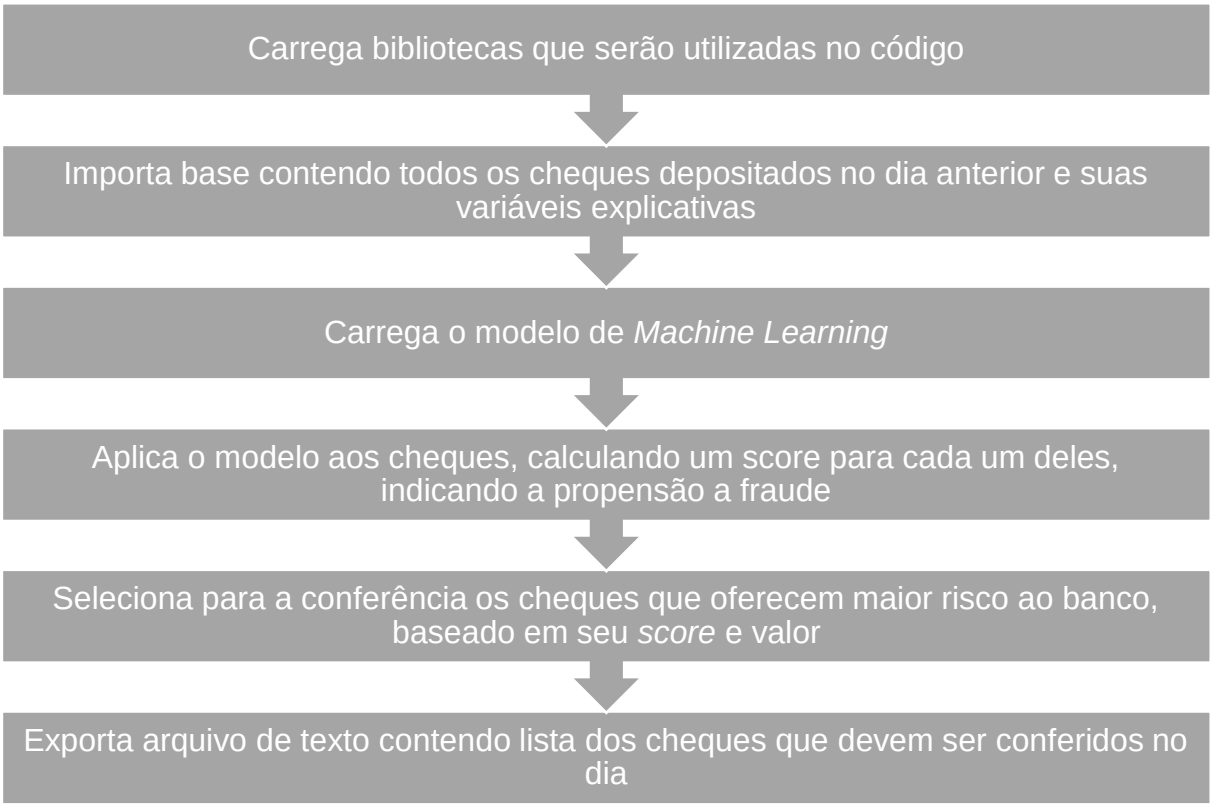

Figura 24 - Estrutura do *script* de aplicação do modelo

(Fonte: elaborado pelo autor)

A Etapa 4 consiste no *upload* do arquivo gerado na etapa anterior no sistema utilizado para seleção dos cheques e sua distribuição entre os operadores que realizarão a conferência.

A sequência de atividades envolvidas nas rotinas de cálculo diária e de aplicação do modelo são apresentadas na Figura 25.

Figura 25 – Sequência de atividades envolvidas na Rotina de Cálculo Diária e na Rotina de Aplicação do Modelo

(Fonte: elaborado pelo autor)

5.4 Rotina de re-treino

É recomendável que o modelo gerado seja atualizado com uma certa frequência. Essa atualização consiste em gerar novas bases de treinamento e teste, para que seja realizada a construção do modelo a partir de um algoritmo pré-definido e aplicado a uma base de testes para verificar sua performance. Os modelos podem se tornar obsoletos com o tempo, e os re-treinos tem como objetivo não permitir uma perda de performance com o passar do tempo. (WEST & BHATTACHARYA, 2014), (POZZOLO, CAELEN, BORGNE, WATERSCHOOT, & BONTEMPI, 2014).

A geração das bases de treino e teste aproveitará o cálculo de variáveis feito diariamente na operação. Após sua geração, serão utilizadas pelo *script* de treino e teste detalhado no Figura 18 na página 52.

As bases de treino e teste serão construídas através de uma rotina no SAS desenvolvida pelo autor. Essa rotina recebe os parâmetros que definem as datas entre as quais os cheques da amostra devem pertencer, parâmetros que definem o volume de dados desejado nas bases de treino e teste e um parâmetro que define o balanceamento da base de treino. A estrutura do programa, exibida na Figura 26, é descrita na Figura 27.

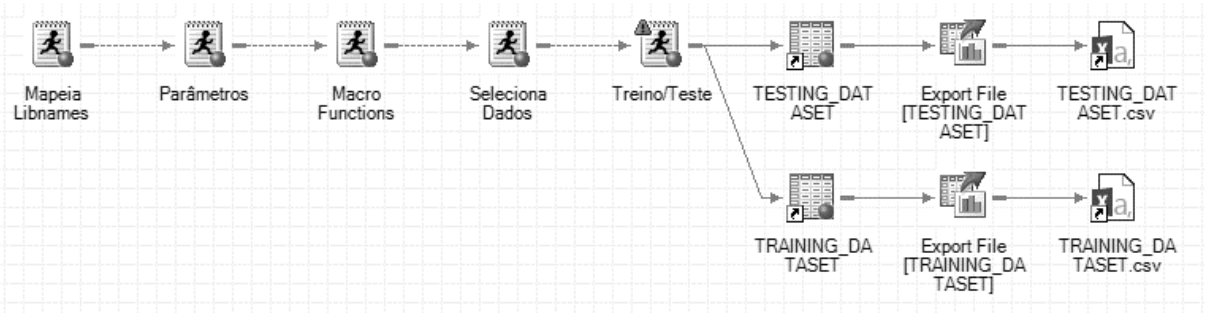

Figura 26 – Programa em SAS para geração de bases de treino e teste

(Fonte: elaborado pelo autor)

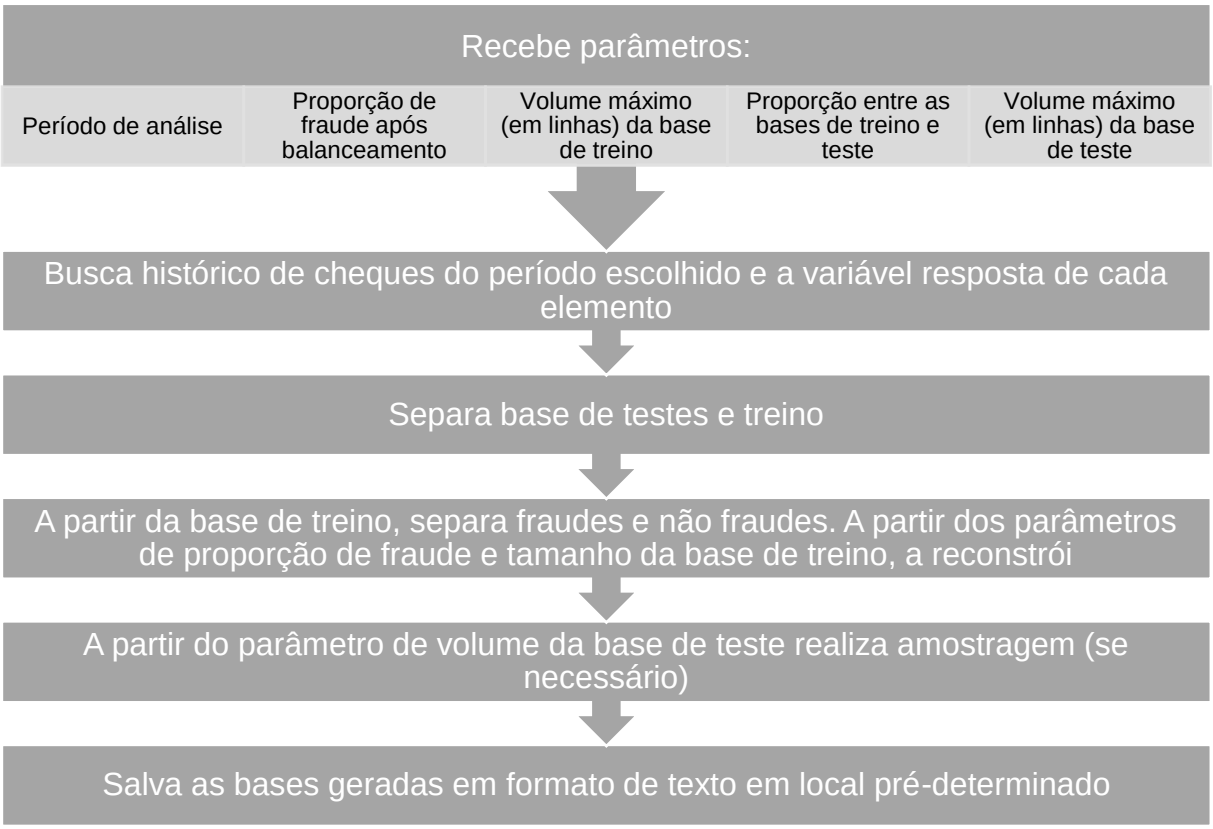

Figura 27 - Estrutura da rotina de geração de bases de treino e teste em SAS

(Fonte: elaborado pelo autor)

6 CONCLUSÃO

Este trabalho teve como objetivo o desenvolvimento de um modelo, baseado em aprendizagem de máquina, capaz de calcular a propensão a ser uma fraude que cada cheque emitido pelos clientes de um banco possui, bem como a implementação deste modelo na área responsável pela conferência de cheques, auxiliando na seleção daqueles mais críticos para análise.

Os resultados obtidos até a conclusão deste trabalho se mostraram satisfatórios, isso porque já foram capturados ganhos tanto na redução das perdas com fraudes quanto na eficiência da operação como um todo, visto que é possível dar maior prioridade aos cheques que de fato apresentem maior risco, reduzindo o tempo total gasto na conferência de cheques.

Para guiar as atividades deste projeto foi escolhida a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) que divide o processo de mineração de dados em seis grandes fases que vão do Entendimento do Negócio à Implantação. A metodologia tem natureza cíclica, se assemelhando ao ciclo PDCA de Deming.

Nota-se que este trabalho foi dividido em duas fases, sendo a Fase I a respeito das cinco etapas iniciais propostas pela metodologia CRISP-DM, enquanto a Fase II tratou exclusivamente da etapa de implantação do modelo desenvolvido na fase inicial.

Ao longo da Fase I foi determinante para o sucesso do projeto a proximidade com o cliente e outros *stakeholders* para um melhor entendimento do processo de compensação de cheques e levantamento das possíveis variáveis que possam explicar a ocorrência da fraude.

Ainda na fase inicial, foi possível estudar e aplicar os princípios e técnicas de mineração de dados que podem ser utilizados para a obtenção de melhores resultados. Além disso foi estudado em maior profundidade o funcionamento e a aplicação de alguns algoritmos de *Machine Learning* e como decisões sobre a parametrização destes modelos podem impactar suas performances.

Ao final do estudo foi verificado que o algoritmo estudado que apresentou maior performance foi o *Random Forest*, que cria diversas árvores de decisão, e utiliza todas elas de forma conjunta para se chegar em uma resposta final.

Também se verificou que utilizar maiores bases de dados resultam em uma performance superior. Outro ponto relevante é que ficou evidente como problemas desbalanceados, em que uma das classes da variável resposta é muito pouco representada, dificultam a aprendizagem dos algoritmos de *Machine Learning*. Para contornar este problema foi utilizada a técnica de

downsampling que trouxe um grande salto de qualidade, sendo a performance máxima observada no balanceamento de 50% (ver Figura 23).

Durante a implantação o grande desafio foi a operacionalização do modelo na rotina da área cliente, ficando evidente como além da performance é necessário planejar com cuidado a viabilidade da aplicação de modelos centrados em grandes volumes de dados. A velocidade e praticidade das atividades são fatores críticos na aceitação do projeto pela área cliente.

Assim como na primeira fase, a proximidade e o bom relacionamento com a área cliente foi essencial para a obtenção dos resultados esperados. Treinamentos com foco nos princípios envolvidos no processo de mineração de dados e algoritmos de *Machine Learning* foram importantes para ter a confiança da área nas sugestões fornecidas pelos algoritmos.

Uma dificuldade do projeto foi o acesso a bases de dados, que nem sempre estavam disponíveis em um mesmo ambiente e bem estruturadas, muitas delas sendo geradas manualmente em MS Excel e MS Access. O investimento na captura, armazenagem e estruturação de dados é essencial para organizações que almejam tirar maior proveito das informações adquiridas em suas operações através de tecnologias emergentes na área de inteligência artificial.

Possíveis avanços do projeto são a automatização dos processos que atualmente são executados manualmente pelos analistas da área e a inclusão de novas variáveis explicativas que não foram identificadas no desenvolvimento deste trabalho.

7 BIBLIOGRAFIA

AGUIAR, S. (2002). *Integração das ferramentas da qualidade ao PDCA e ao Programa Seis Sigma.* Belo Horizonte: Editora DG.

ANDRADE, F. F. (2003). *O Método de Melhorias PDCA.* São Paulo.

AZEVEDO, A., & SANTOS, M. F. (2008). KDD, SEMMA and CRISP-DM: A Parallel Overview. *IADIS European Conference Data Mining.* Amsterdam.

BADIRU, A., & AYENI, B. (1993). *Practitioner's Guide to Quality and Process Improvement.* Londres.

Banco Central do Brasil. (2014). Fonte: http://www.bcb.gov.br/pre/bc_atende/port/servicos6.asp?idpai

Banco Central do Brasil. (2014). Fonte: http://www.bcb.gov.br/pre/bc_atende/port/servicos7.asp

BHATTACHARYYA, S., JHA, S., THARAKUNNEL, K., & WESTLAND, J. C. (2011). Data mining or credit card fraud: A comparative study. *Decision Support Systems*, pp. 602-613.

BREUR, T. (20 de Agosto de 2017). *Deming's PDCA Cycle & Data Science*. Fonte: Data, Analytics and beyond: https://tombreur.wordpress.com/2017/08/20/demings-pdca-cycle-data-science/

CARVALHO, M. M., & ROTONDARO, R. (2002). *Seis Sigma: estratégia gerencial para melhoria de processos, produtos e serviços.* São Paulo: Atlas.

Cornell University Law School. (2009). *White-Collar Crime: an overview.* Fonte: https://www.law.cornell.edu/wex/white-collar_crime

GOLDEN, T. W., SALAK, S. L., & CLAYTON, M. M. (2006). *A Guide to Forensic Accounting Investigation.*

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2001). *The Elements of Statistical Learning.*

ISTOÉ. (2016). Cheque fraudado: quem paga a conta?

KDnuggets. (Outubro de 2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. Fonte: KDnuggets: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

MARCONDES, J. S. (2017). Conceito de Fraude - O que é? Definição, Significado, Prevenção.

MARISCAL, G., MARBÁN, O., & COVADONGA, F. (Junho de 2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, 25*, pp. 137-166.

NGAI, E., HU, Y., WONG, Y., CHEN, Y., & SUN, X. (2010). The application of data mining techniques in financial fraud detection: A classification. *Decision Support Systems*, pp. 559-569.

OLIVEIRA, R. L. (2012). *Gestão de fraudes financeiras externas em bancos.* Ribeirão Preto.

PHUA, C., LEE, V., & GAYLER, K. S. (2005). A comprehensive survey of data mining-based. *Artificial Intelligence Review*, pp. 1–14.

POZZOLO, A. D., CAELEN, O., BORGNE, Y.-A. L., WATERSCHOOT, S., & BONTEMPI, G. (2014). Learned lessons in credit card fraud detection from practitioner perspective. *Expert Systems with Applications*, pp. 4915-4928.

PROVOST, F., & FAWCETT, T. (2013). *Data Science for Business.*

TERNER, G. L. (2008). *Avaliação da aplicação dos métodos de análise e solução de problemas em uma empresa metal-mecânica.* Porto Alegre.

TURBAN, E., ARONSON, J., LIANG, T., & SHARDA, R. (2007). *Decision Support and Business.* Pearson Education.

WEST, J., & BHATTACHARYA, M. (2014). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*.

WEST, J., & BHATTACHARYA, M. (2016). Some Experimental Issues in financial Fraud Mining. *Procedia Computer Science*.