Uma Abordagem Bayesiana para Análise de Fraude de Subscrição em Telecomunicações

Elizabeth Agnes Urban Cristofaro

Orientador: Prof. Dr. Francisco Louzada-Neto

Dissertação a ser apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos

Dezembro- 2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C933ab

Cristofaro, Elizabeth Agnes Urban.
Uma abordagem bayesiana para análise de fraude de subscrição em telecomunicações / Elizabeth Agnes Urban Cristofaro. -- São Carlos : UFSCar, 2011.
104 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Estatística. 2. Fraude. 3. Telefonia celular. 4. Inferência bayesiana. 5. Regressão logística. I. Título.

CDD: 519.5 (20ª)

"Esta manhã, antes do alvorecer, subi numa colina para admirar o céu povoado, E disse à minha alma: Quando abarcarmos esses mundos e o conhecimento e o prazer que encerram, estaremos finalmente fartos e satisfeitos?
E minha alma disse: Não, uma vez alcançados esses mundos prosseguiremos no caminho."

Walt Whitman

Dedico este trabalho ao meu companheiro e cúmplice Francisco, por sempre ter acreditado em mim, até nos momentos em que já havia esmorecido. À meu filho Fernando, por sempre me impulsionar a me aprimorar. Amo muito vocês! (Não posso esquecer de um ser muito especial - meu muito obrigada!)

Aos meus colegas do Mestrado em Estatística UFSCAR, pelos momentos de convívio, risos, trocas e afetos. Com muita saudade, obrigada.

Ao pessoal da Secretaria Acadêmica, pela eficiência, dedicação e simpatia. Meu muito obrigada.

A todos os professores que participaram desta jornada, sempre solícitos, até mesmo fora do horário do curso, porque sem eles não haveriam sonhos e idéias. Meus sinceros agradecimentos.

Ao meu orientador, que se portou como um mestre. Acreditando no meu trabalho, deu-me a liberdade necessária dividindo comigo as expectativas, conduziu-me a maiores reflexões e desta forma enriqueceu-o. Minha especial admiração e gratidão.

Resumo

Este trabalho tem por objetivo comparar a performance da inferência bayesiana e inferência clássica na classificação de comportamento do fraudador (considerado evento raro). Espera-se desenvolver um método para inferir e internalizar novos padrões de fraude baseado na abordagem bayesiana, possibilitando a construção do conhecimento sobre o evento a partir da inclusão de informações históricas incrementais em funções encadeadas.

Palavras-chave:.fraude de subscrição, celular, inferência bayesiana, regressão logística

Sumário

Capítulo 1

Introdução

A fraude em Telecomunicações tem causado prejuízos de bilhões de dólares às operadoras em todo o mundo. Estima-se que a perda anual por fraude, na industria de Telecomunicações, em todo o mundo, supera os 35-40 bilhões de dólares.[1]

A fraude continua restringindo os ganhos da indústria de Telecomunicações dos EUA, o uso fraudulento das redes e o roubo de serviços, em todos setores da indústria, crescem de 10 % a 12% ao ano. Estima-se que a perda anual para a fraude de subscrição está entre 4 e 22 bilhões dólares, chegando, em alguns casos, em até 10% do balanço de uma operadora.[2]

Conseqüentemente, as perdas de receita são enormes devido ao não pagamento das chamadas realizadas de má fé. No entanto, o custo dessas ações não são facilmente quantificadas, devido a complexidade do dimensionamento, perda de integridade da rede e a redução da satisfação do cliente, entre outros fatores.

Diariamente, inúmeros tipos de fraudes são registrados pelas empresas de Telecomunicações, sendo as fraudes de voz (calling fraud) as responsáveis pelos maiores prejuízos financeiros registrados pelas operadoras.

O desenvolvimento de novos serviços de valor agregado, abertura de mercado e aumento da competitividade, tem impulsionado as operadoras ao aumento da carteira de clientes através da adesão de novos clientes e a oferta de serviços diferenciados.

A fraude em Telecomunicações tem representado enormes perdas de receita para as

[1] 2003 Fraud Loss Press Release, www.cfca.org

[2] Guerra, J. 2004 *Telecom Fraud on the Rise* in www.billingworld.com

operadoras, sendo necessário o desenvolvimento de um sistema de gerenciamento com métodos eficazes e abrangentes no combate à ação dos fraudadores. O desafio se torna ainda maior com o surgimento de modernas tecnologias no setor, em constante evolução, assim como os meios encontrados pelos fraudadores em burlar os sistemas de segurança desenvolvidos pelas operadoras.

Os custos reais da fraude podem ser muito superiores ao da receita perdida; como por exemplo: desvio de recursos, investimentos desnecessários na rede, perda de clientes. O impacto real raramente é quantificado, o que reduz a visibilidade e a eficiência de medidas e ações de combate à fraude.

Qualquer investigador deparado com a necessidade da análise de dados, precisa de fazer uma escolha racional sobre o método particular de análise. Devem ser levadas em conta algumas considerações importantes nessa escolha, como por exemplo:

- o objetivo da investigação;
- as características matemáticas das variáveis envolvidas;
- as hipóteses estatísticas feitas sobre estas variáveis;
- como foram recolhidos os dados;
- a natureza do evento em estudo.

As duas primeiras considerações são, de um modo geral, suficientes para determinar uma análise apropriada. No entanto, o investigador deve também considerar os outros itens antes de finalizar a recomendação.

Para certos acontecimentos, como no caso da fraude, espera-se desenvolver modelos preditivos baseados em observações de determinado fenômeno que permitam a previsão ou detecção desse acontecimento numa fase incipiente de desenvolvimento.

Outra questão que se torna problemática na modelagem preditiva está associada à exatidão e precisão. A precisão está associada à dispersão dos valores em sucessivas observações, enquanto que a exatidão refere-se à proximidade de uma estimativa do verdadeiro valor que pretende representar. As limitações da exatidão e da precisão originaram os conceitos de sensibilidade e especificidade de um teste de diagnóstico. Estas medidas e os índices a elas associados, como a proporção de verdadeiros positivos e a proporção de falsos positivos, são mais significantes do que a exatidão, embora não forneçam uma descrição única do desempenho do modelo.

O maior problema da sensibilidade e da especificidade é que estas medidas dependem do critério de diagnóstico ou de um valor de corte, o qual é por vezes selecionado arbitrariamente. Assim, mudando o critério pode-se aumentar a sensibilidade com o consequente detrimento da especificidade, e vice-versa. Deve-se considerar também que um critério de decisão particular depende também dos benefícios associados aos resultados corretos e dos custos associados aos incorretos.

1.1 Motivação

O Gerenciamento de Fraudes tem por objetivo detectar, prevenir e minimizar eventos de fraude em Operadoras , através do desenvolvimento de modelos preditivos capazes de identificar corretamente padrões suspeitos.

O objetivo é que todo conhecimento sobre as ações fraudulentas seja gerado e transferido em tempo real ou quase real. É neste ponto que esta linha de pesquisa procura apoio na inferência bayesiana pois, como os padrões de fraude são altamente voláteis, os modelos embasados nas técnicas estatísticas clássicas e de inteligência artificial pressupõe uma massa considerável de informações históricas.

Capítulo 2

A Fraude

Uma definição suscinta de fraude é: "O uso ilícito de acesso a rede de telefônia celular para obter proveito ou lucro". Os fraudadores são muitos bem organizados e exploram tipicamente o ponto mais fraco disponível na operadora de telefonia celular, ou seja, procuram sempre a maneira mais fácil e mais barata de obter vantagem da operadora.

A fraude de telefonia é um fenômeno mundial. Estimativas correntes contabilizam perdas de US$ 15 a U$ 55 bilhões por ano (1% a 5%) na indústria de telefonia, que movimenta negócios na ordem de US$ 1,5 trilhão. Vale ressaltar que estas estimativas estão relacionadas ao que pode ser medido, ou seja, as fraudes conhecidas, e que provavelmente os valores podem ser ainda mais elevados devido aos casos de perdas que não são contabilizadas como fraude.

A Fraude é um negócio rentável, estável e muito bem organizado e administrado.

2.1 Objetivo do Fraudador

Os objetivos principais do Fraudador são:

- Obtenção de lucros.
- Esconder sua identidade através de um assinante real, funcionando como um disfarce ideal para o tráfico de drogas, o crime organizado, o terrorismo e a lavagem de dinheiro. O fraudador simplesmente não quer que suas ligações sejam rastreadas e que permitam que ele seja descoberto.

- O desafio, para provar que pode ser feito, como é o caso dos "hackers".

2.2 Tipos de Fraudes

Os principais tipos de Fraudes são descritos a seguir.

2.2.1 Fraudes Técnicas

Clonagem

É a “cópia” desautorizada da identidade de um terminal para permitir que as chamadas sejam cobradas de um cliente válido. Neste cenário, os números de identificação de celulares válidos (MIN) e os números de série eletrônicos (ESN) ou IMSI (redes GSM) são obtidos na rede de telefonia celular. Estas combinações válidas de MIN/ESN podem ser adquiridas de várias maneiras, entre elas pela “escuta” nos canais do controle e de comunicação dos terminais, ou através de uma fonte interna da operadora de telefonia celular.

Auto-Clonagem

Os assinantes clonam seus próprios telefones para criar uma extensão. Pretendem pagar a conta (e a maioria paga), mas é ilegal, e a intenção é a de não pagar as taxas extras cobradas pela operadora para telefones e/ou assinaturas adicionais.

Telefone Mágico

É um dispositivo do que pode ser programado com pelo menos 100 MINs e ESNs e muitas chamadas podem ser feitas em seqüência, usando um MIN - ESN diferente em cada vez. Isto espalha a fraude entre muitos assinantes, impedindo um aumento rápido do uso de um mesmo telefone. Isto faz com que seja mais difícil e mais demorado detectar a fraude.

Seqüestro (Hijacking)

Utiliza-se de transmissores de alta potência para capturar o sinal do telefone do assinante, assim que o processo de autenticação é completado. O transmissor pirata realiza, então,

as chamadas utilizando-se da característica do serviço "siga-me" (call forwarding). O assinante "real" é desconectado da chamada e não sabe que o seu telefone ainda está sendo usado.

2.2.2 Fraude de Roaming

É a Fraude na qual obtêm-se telefones celulares ilegalmente ou cartões SIM adulterados (GSM) para fazer chamadas na operadora visitada. O tempo de atraso de envio de registros de chamada incentiva e aumenta seus efeitos.

2.2.3 Fraude em Serviços de Valor Agregado (PRS-Premium Service)

Esta Fraude consiste em aumentar desonestamente o valor devido a um fornecedor de serviço (fraudador), organizando chamadas para esse serviço. Os fraudadores ativam o serviço em seu nome e normalmente usam cartões Pré-Pagos (para se manterem anônimos) efetuando as chamadas para o número de seu serviço, as quais são faturadas de forma incorreta, ou seja, o valor da ligação é muito menor do que a tarifa do serviço paga pela operadora ao proprietário do serviço (fraudador).

Como no Brasil atualmente efetua-se o cadastramento dos usuários de terminais pré-pagos, este tipo de fraude torna-se mais difícil e perigosa de ser praticada.

2.2.4 Fraude de Revenda

A Fraude de Revenda pode ou não incluir a participação de um revendedor. Neste caso, as solicitações para novas linhas telefônicas são aceitas e aprovadas sem verificação apropriada da identidade ou sem informação suficiente o bastante para produzir uma conta para o assinante. O revendedor pode não ser desencorajado de tal ação pelo fato que freqüentemente são pagos pelo número e não pela qualidade das ativações. Uma maneira popular, e em ascensão, é a do revendedor aumentar suas comissões ativando assinantes inexistentes, que morreram, que se mudaram, e etc.

Estes assinantes novos não pagam as suas contas, e também não fazem chamadas. Após um determinado período estes assinantes são desativados e para a operadora tudo

se passa com se não houvesse nenhum prejuízo, já que não houve nenhuma chamada a ser cobrada. Entretanto, o revendedor ganhou sua comissão.

2.2.5 Fraude de Engenharia Social ("Social Engineering")

A Fraude de Engenharia Social ocorre quando um fraudador convence um empregado da operadora a revelar a informação necessária para cometer a fraude, ou quando convence o empregado a cometer a fraude ele mesmo. Isto se dá sem que a vítima saiba que está ajudando o fraudador. O problema com os bons fraudadores é que são muito convincentes, e não desistem até que conseguem o que querem. Inicialmente as pessoas questionam, mas eventualmente fazem o que o fraudador quer. Saber algumas informações conhecidas (nome do chefe ou coordenador), ajuda a estabelecer a credibilidade do fraudador.

2.2.6 Fraude Interna

A Fraude Interna é cometida quando um empregado da operadora ajuda na obtenção de informações, serviços opcionais, ou equipamentos que permitem que o fraudador obtenha o acesso à rede ou ao serviço sem pagar por ele.

Alguns indicadores de fraude interna são:

- Distribuição de MINs e de ESNs de celulares existentes, ou chaves de autenticação;
- A criação de assinantes que não são faturados ou cobrados (telefones fantasmas), isto é, um assinante é ativado na rede, mas o sistema de faturamento deliberadamente não recebe esta informação;
- Uso abusivo das linhas de testes ou de emergência. Um aumento significativo de Fraudes Internas pode ocorrer quando estas linhas não são controladas rigidamente e também não são faturadas.

2.2.7 Fraude de Subscrição

A Fraude de Subscrição se caracteriza pela apresentação de informação imprecisa ou incorreta para obter um contrato de serviço ou, por outro lado, pelo não cumprimento das obrigações desse contrato. A Fraude de Subscrição ocorre quando um assinante contrata

o serviço com identificação falsa ou informação fraudulenta obtida de cliente “real”, e não tem nenhuma intenção de pagar pelo serviço. Difere da inadimplência no sentido que é um ato deliberado para roubar o rendimento da operadora, ao contrário do inadimplente. Geralmente a fraude de subscrição fica escondida entre as perdas de inadimplência, ou seja, a operadora não a identifica como fraude e nunca recupera a perda.

Um dos meios mais comuns (e mais eficazes) de cometer a fraude de subscrição é o roubo de informações de identidade. Isto ocorre quando as informações do assinante são usadas para obter serviços de forma ilegal por um fraudador (para compra de bens, abrir contas em banco, tomar empréstimos, etc.). Estes atos são realizados sem o conhecimento do assinante cuja identidade foi usada (a vítima) e pode levar a uma situação de histórico de crédito ruim e a outros sérios incômodos. Como um exemplo, hoje em dia é possível comprar um celular ou um serviço por telefone fornecendo somente nome, número da identidade e alguns dados pessoais, e o endereço. Naturalmente tudo isso pode ser falsificado ou o fraudador pode se fazer passar por outra pessoa "responsável"pela compra.

Fraude de Aparelho & Aparelhos Subsidiados

Outra forma de Fraude de Subscrição em redes celulares vem dos negócios paralelos estabelecidos por fraudadores:

- Reciclagem de aparelhos roubados;
- Revenda de aparelhos subsidiados;
- Aparelhos forjados.

Este tipo de fraude é detectado como parte da detecção de clonagem e de fraude de subscrição. Quando disponíveis, as lista negras dos aparelhos roubados são também utilizadas no sistema de detecção de fraude, a fim de gerar alarmes sempre que um aparelho dessa lista estiver em uso.

Fraude Pré-Pago

Ao contrário do que muitos pensam, o serviço de Pré-Pago em redes celulares também está exposto à fraude. Algumas dessas fraudes são:

- Efetuar recarga fraudulenta de créditos;
- Impedir a dedução do saldo da chamada;
- Invasão ("hacking") de certos tipos de aparelhos para parar de deduzir a chamada atual do saldo.

Capítulo 3

Gerenciamento da Fraude

A avidez para ganhar a participação no mercado e apressar-se em introduzir novos produtos, faz com que as exigências e regras de análise de crédito para novos assinantes sejam reduzidas, conseqüentemente aumentando a exposição à fraude.

Os Fraudadores freqüentemente são os primeiros a experimentar novos produtos e serviços, em busca de vulnerabilidades existentes para cometer a fraude.

3.1 Combate à Fraude

O combate à Fraude compreende a redução das perdas da operadora a um nível pré-estabelecido. A eliminação da fraude não é viável, pois seria mais caro eliminá-la do que o próprio custo da Fraude. Os Sistemas Antifraude podem reduzir as perdas da fraude, mas para minimizá-la realmente, as operadoras devem aplicar técnicas de prevenção à fraude por todo os processos da rede e do negócio. Deve-se ter em mente que o nível aceito está relacionado ao que pode ser mensurado, ou seja, a fraude já conhecida.

3.2 Objetivo do Combate à Fraude

O objetivo é minimizar as reclamações, do assinante "verdadeiro", relativas ao recebimento de contas altas, a mudança do número do celular ou a existência de chamadas desconectadas, de forma a reduzir o "churning" (mudança de operadora) e o descontentamento dos clientes. O assinante "verdadeiro"não pagará pelo uso desautorizado, mas a

operadora ainda terá que pagar pelo uso de interconexão, ou seja, pelo uso eventual da rede das outras operadoras para completar a chamada.

3.3 Ciclo de Vida da Gerência de Fraude

O Ciclo de Vida da Gerência de Fraude é dinâmico e adaptativo. Há oito estágios do ciclo de vida: Intimidação, Prevenção, Detecção, Medidas para parar a Fraude, Análise, Política, Investigação e Acusação.

- Primeiro Estágio: Intimidação

É caracterizado pelas ações e pelas atividades destinadas a inibir ou desanimar o fraudador antes de executar a fraude, por medo das conseqüências. O possível fraudador não tentará porque as probabilidades do sucesso da fraude não são suficientemente boas.

- Segundo Estágio: Prevenção

Dado que o estágio de intimidação não dissuadiu o fraudador de cometer a fraude, este estágio pretende impedir os fraudadores de ter sucesso. Compreende atividades que tornam a execução da fraude mais difícil, endurecendo as "defesas"contra os fraudadores. Exemplo: Autenticação do aparelho.

- Terceiro Estágio: Detecção

Conjunto de ações e de atividades, tais como sistemas antifraude, são utilizadas para identificar e encontrar a fraude antes, durante e depois da conclusão da atividade fraudulenta. A intenção da detecção é descobrir ou revelar a presença da fraude ou de uma tentativa de fraude.

- Quarto Estágio: Medidas

O objetivo é a tomada de medidas que evitem a ocorrência de perdas ou a sua continuidade, e ou impeçam um fraudador de continuar a fraudar ou terminar a sua atividade de fraude. Exemplo: executar a desativação de um assinante.

- Quinto Estágio: Análise

Perdas que ocorreram apesar dos estágios precedentes, neste estágio são identificadas e estudadas para determinar os fatores em que ocorreram as fraudes.

- Sexto Estágio: Política

Compreende o conjunto de atividades que pretendem criar, avaliar, comunicar e ajudar na implantação de políticas para reduzir a incidência da fraude. Equilibrar as políticas de redução de fraude com as restrições de orçamento e gerência eficaz é uma obrigação neste estágio.

- Sétimo Estágio: Investigação

A investigação envolve obter evidências suficientes e informações para parar a atividade fraudulenta, para recuperar recursos ou obter a restituição dos mesmos.

- Oitavo Estágio: Acusação

A Acusação bem sucedida e a condenação do fraudador dependem na maior parte do estágio precedente. Neste estágio está clara a necessidade do suporte jurídico (leis) para condenar os criminosos.

Através de uma interação integrada e equilibrada entre os estágios, podem ser alcançados resultados mais eficazes e mais eficientes. Cada operadora deve descobrir o melhor equilíbrio e interação para o seu negócio. Estes estágios não são necessariamente executados na ordem apresentada, e interagem com todos os outros estágios restantes dinamicamente. Devem estar espalhados por todos os departamentos da operadora, reforçando a idéia de que a fraude não é somente uma preocupação do departamento de fraude e sim da operadora como um todo.

Conseqüentemente, canais de comunicação formais devem ser estabelecidos de maneira a facilitar o fluxo da informação entre as áreas envolvidas, realçando a cooperação em vez da competição destrutiva entre os departamentos. As operadoras normalmente focam em torno da atividade fraudulenta ou do fraudador e não em torno da gerência e da redução de perdas com a fraude.

Vale ressaltar que o compartilhamento de informações de fraude entre as operadoras, a fim de responder rapidamente a novas ameaças torna-se bastante importante. Como concorrentes estas informações raramente são trocadas.

3.4 Sistemas Antifraude

Os Sistemas Antifraude se encaixam na maior parte no estágio da detecção. Os dados obtidos pelos sistemas também ajudam no estágio da análise, na investigação e acusação.

Geralmente, os sistemas antifraude monitoram a utilização (“uso") da rede celular, para detectar padrões da fraude, como observamos na figura 3.4.

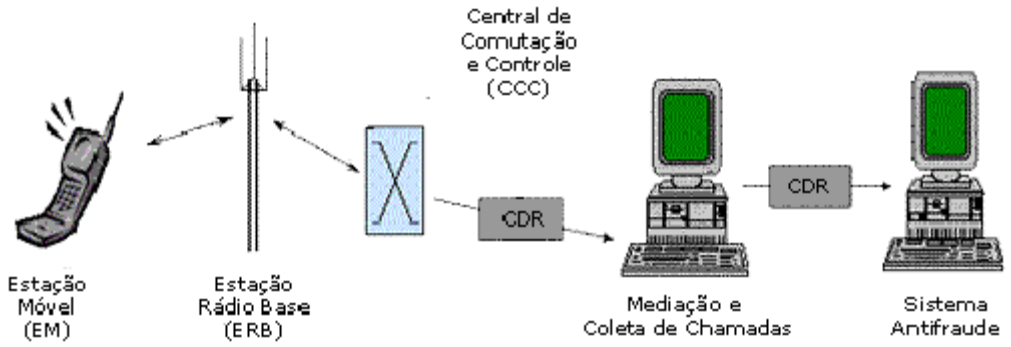


Figura 3.1: Arquitetura Sistema Antifraude

Quando uma chamada é realizada e completada (terminada) com sucesso e o assinante desliga o telefone celular, a ERB com a qual a EM estava se comunicando finaliza a conexão e a CCC grava o registro de sua chamada gerando um CDR ("Call Detail Record"). O sistema de mediação então coleta estes CDRs e os entrega a vários sistemas, entre eles o sistema antifraude.

Este tipo de análise é chamado de Pós-evento porque os CDRs são gravados somente depois que a chamada é completada ou terminada. Quanto mais cedo este CDRs alcancem o sistema antifraude, mais rapidamente a detecção poderá ser realizada.

Tipicamente, toda a atividade incomum ou suspeita resulta em um alarme. A análise destes alarmes, baseado em conhecimentos pré-configurados e adquirido pelo sistema e no padrão de comportamento do assinante, geram casos de fraude. Estes casos são então enviados aos analistas de fraude para análise e conclusão e podem assumir os seguintes resultados: Fraude, Não Fraude e Desconhecido.

Um conceito muito importante é aquele relativo aos casos de Fraudes Falso-positivas. O sistema cria um caso de fraude e o analista decide que não é fraude. Por mais que o

sistema tenha informações sobre o padrão de comportamento do assinante e sua base de conhecimento esteja corretamente configurada, ainda assim podem existir situações que não se configuram em casos de fraude, após uma análise pormenorizada do analista de fraude.

Os sistemas antifraude tratam de volumes muito grandes de dados e são muito sensíveis ao desempenho. Cada vez mais estão recebendo dados de outras fontes de informação para não ficar limitados a técnicas da detecção em função do "uso". Essas informações adicionais podem ser: Informação de tarifação (Billing), Informações do assinante e informações de associações de proteção ao crédito (Exemplo: Serasa);

Os diversos relatórios e a definição de KPI (indicadores chaves do desempenho) são usados para monitorar o desempenho do sistema antifraude.

3.4.1 Técnicas de Detecção de Fraude

Essas técnicas podem ser divididas em dois grupos, os quais são apresentados a seguir.

Técnicas Estáticas

As técnicas estáticas de detecção são:

- Detecção da colisão (sobreposição da chamada): determina se duas chamadas, supostamente do mesmo telefone ou serviço, ocorreram ao mesmo tempo.
- Verificação (geográfica) da velocidade: determina se duas chamadas, supostamente do mesmo telefone ou serviço, ao serem feitas em duas posições geográficas suficientemente distantes e num período muito curto de tempo, são passíveis de acontecerem.
- Desvio de Perfil de Uso: detecta quando um perfil individual recentemente calculado para um dado assinante difere de seu perfil precedente por mais do que uma quantidade especificada pela operadora.
- Lista Negra: para cada chamada que chega ao sistema, são verificadas as listas de aparelhos roubados, MIN/ESN, e de IMSI que podem ser usados para a clonagem. Quando existe coincidência de informações, um caso de fraude é gerado para análise.

Técnicas Dinâmicas

Estas técnicas se destinam a descobrir novos padrões e novas tendências de fraude, e são:

- Detecção de Padrões (Criação de Regras): Incrementa a experiência ou a intuição do Analista de Fraude com a aquisição contínua de conhecimento e de treinamento, o que permite a percepção de novos métodos de fraude e a definição das regras e padrões para encontrar mais fácil e rapidamente as fraudes.

- Rastreamento de Chamadas: esta técnica executa a análise detalhada das chamadas recebidas e originadas de/para assinantes suspeitos de fraude. A título de exemplo, um traficante de drogas muito provavelmente permanecerá em contato constante com outros criminosos (comparsas) através de aparelhos celulares. Através desta técnica, torna-se possível a visualização de todas as chamadas recebidas e originadas pelo assinante suspeito. É possível ainda, de modo gráfico, visualizar o encadeamento em vários níveis de um número desejado. Esta técnica é muito útil para descobrir prováveis novos fraudadores ou para a investigação policial.

- Repositório de Dados e Mineração de Dados ("Data Warehouse and Data Mining"): oferecem técnicas avançadas de análise através de métodos estatísticos e de inteligência artificial e de refinamentos sucessivos, a partir de dados de alto nível descendo a níveis de detalhes cada vez maiores para uma análise interativa. Através destas técnicas pode-se chegar a descoberta de novos padrões de fraude e a fraudes existentes ainda desconhecidas.

- Pontuação baseada em Redes Neurais: para cada novo caso de fraude, o sistema calcula uma pontuação como sendo a similaridade com os casos conhecidos, em que os analistas de fraude decidiram ser fraude, como os casos não fraude. Permite ao analista maior confiança para decidir se é ou não realmente fraude.

- Assinatura (Impressão Digital) do assinante: cria uma chave que identifica unicamente cada assinante por uma "assinatura"calculada baseado no "uso"e nos dados de tarifação. Um repositório de assinaturas é criado e é utilizado para verificar se novos assinantes não são fraudadores reincidentes.

3.5 Cenários e Regras utilizadas para detecção de Fraude em Telecomunicações

A fraude consiste na utilização de meios ilícitos para obtenção de vantagens para o fraudador e que se reverte em prejuízo para as operadoras, à medida que serviços são utilizados e não são pagos ou, então, são cobrados à pessoa errada.

Existem vários tipos de fraude e muitas regras já foram identificadas como características de fraude. Mesmo assim, os fraudadores procuram inovar, idealizando novas ações de má fé para atingirem seus objetivos em detrimento da operadora de Telecomunicações. Como são diversos os tipos de fraude também serão distintas as técnicas utilizadas para se detectar os inúmeros padrões existentes. Os modelos estatísticos utilizados são construídos a partir de técnicas estatísticas e inteligência artificial (redes neurais), onde se pode determinar as relações entre a fraude e as variáveis identificadas.

3.5.1 Fraude de Subscrição

Atualmente, este é o tipo de fraude predominante nas operadoras, em que a pessoa faz a habilitação usando dados cadastrais falsos. Com o aumento da segurança das plataformas de terceira geração nas celulares (GSM), a fraude de subscrição passou a ser a principal ofensora.

Cenário

A diversificação do canal de vendas, tornando mais difícil a validação das informações do usuário e a instalação / habilitação do serviço antes de emitir uma fatura, podendo usar até cortar a linha, são aspectos que motivam a fraude de subscrição.

Regras

No momento do cadastro de novos assinantes deve-se verificar a autenticidade da documentação fornecida. Para isso é preciso dispor de sistemas que façam checagens automáticas, como validação de CPF's, conferências de endereço e CEP, informações de óbito, etc. Outro interesse nesse momento são as famosas listas negras, onde se pode

buscar a ocorrência de registros que comprometam a idoneidade desses clientes, como má fé na sustação de cheques, entre outros. No caso de clientes jurídicos, devem-se buscar outras informações, como relacionamento com outros fornecedores.

Num segundo momento é importante investigar as alterações das informações do assinante logo após a ativação do sistema, o que pode ser um indicativo de atitude de má fé do novo cliente.

A empresa operadora deve analisar de perto a utilização e o comportamento dos novos assinantes que adentram á sua base, acompanhando e comparando as características apresentadas pelos mesmos com o perfil dos seus clientes normais. Normalmente as operadoras costumam classificar o cliente como suspeito de fraude nos 3 (três) primeiros meses de relacionamento

3.5.2 Fraude Técnica

Apesar dos casos de clonagem de celulares representarem atualmente pequena proporção do total das fraudes nas operadoras de telefonia móvel, o prejuízo é inversamente proporcional, pois a clonagem gera, em sua maioria, diversas ligações de longa distância, principalmente internacionais, as quais são mais difíceis de serem controladas.

A duplicação de linhas celulares já gerou e ainda gera demasiados problemas para as operadoras, quando analógicas, e imaginaram estarem livres deste risco com o desenvolvimento do celular digital. Entretanto, as técnicas dos fraudadores evoluem em paralelo as desenvolvidas no sentido de fornecer segurança contra eles. O fraudador copia o número de série da identificação de um celular habilitado no momento em que o assinante se comunica com a operadora para efetuar uma chamada telefônica ou, consegue a combinação de ESN (*Eletronic Serial Number*) e número do acesso de uma outra maneira.

A clonagem de celular é a forma mais comum de se entender uma fraude técnica. Muitos outros nomes são identificados no mercado de fraude, como forma de diferenciar as diversas técnicas e métodos de gerar uma fraude técnica tais como: tumbling, surfing, black boxes, blue boxes, clip-on, etc.

Na fraude caracterizada como Tumbling, um celular reprograma aleatoriamente seu número telefônico ou Número Serial Eletrônico (ESN) após cada chamada, tirando proveito do processo de validação de chamadas que ocorre quando um usuário realiza sua primeira

chamada fora da área para qual o telefone foi habilitado.

- *Surfing* é caracterizado como o uso do serviço de terceiros sem autorização.
- *Black Boxes* é a conexão de um dispositivo na rede para simular os sinais para conseguir uma linha para fazer a ligação. Ninguém paga pela ligação.
- *Blue Boxes* é a conexão de um dispositivo na rede para usar a linha de outra pessoa. O dono da linha é quem paga a conta.
- *Clip-on* é a utilização de grampo na linha telefônica de outra pessoa. Quem paga a conta é o dono da linha.

Cenário

Com o avanço tecnológico, hoje já se sabe que é possível clonar um celular com tecnologia digital TDMA (*Time Division Multiple Access*) e CDMA (*Code Division Multiple Access*), embora com mais dificuldade em relação àqueles de tecnologia analógica. Já o GSM (*Global System for Mobile Communications*) possui um chip criptografado e é considerado pelos analistas como o mais seguro, entretanto sabe-se que também é possível a clonagem de um celular GSM com o auxílio de algoritmos que decodificam essas criptografias.

A investigação desta fraude é basicamente sobre aspectos geográficos, mas também pode ser sinalizada quando verificadas mudanças bruscas nos padrões de comportamento dos clientes.

A identificação da fraude técnica deve ser feita quanto antes, visando diminuir o valor da perda. Também é muito importante que as vitimas não sejam surpreendidas com faturas de valores absurdamente elevados, ou seja, a rapidez da identificação da fraude deve estar associada ao processo de estorno destas ligações da fatura do cliente antes de rodar seu ciclo de faturamento.

Regras

Uma clonagem é imediatamente detectada quando se observa que há a sobreposição de chamadas, ou seja, duas chamadas sendo realizadas ao mesmo tempo a partir do mesmo número telefônico, sem que esse disponha do serviço de chamada em espera.

A mesma filosofia se dá ao investigar os locais e horários de onde se origina a ligação. Por exemplo: é impossível que após uma ligação gerada em São Paulo às 10 horas, se dê uma outra em Salvador às 11 horas do mesmo dia.

Um alarme deve ser acionado quando se percebe mudanças bruscas no comportamento do cliente, como aumento nas ocorrências de ligações internacionais, aumento no volume das chamadas nacionais, uso de serviços de conferência, transferências de chamadas e etc.

Algumas fraudes técnicas estão sendo identificadas somente quando o cliente recebe a sua conta com um valor não usual, então a empresa operadora deve normalmente analisar a utilização e o comportamento de todos seus assinantes, acompanhando e comparando as características de uso ao seu perfil histórico. Todos os clientes, sem exceção, devem ser analisados. Por exemplo: Um estudante que tem um gasto mensal de R$ 60,00 (sessenta Reais) não pode receber uma fatura cobrando R$ 12.452,12 (Doze mil, quatrocentos e cinqüenta e dois reais e doze centavos) pelos serviços que foram prestados.

3.5.3 Fraude Interna

Muitas vezes, as atitudes fraudulentas que atacam as operadoras podem ser decorrentes de pessoas que estão lá dentro e têm acesso ás informações confidenciais ou com capacidades para habilitação de serviços.

Cenário

O cenário desta fraude é composto de funcionários que vendem informações, ativam telefones fantasmas ou serviços adicionais em uma linha sem cobrar ou alteram bilhetes e faturas. Podem também ocultar informações que identifiquem uma fraude e não verificar as ligações rejeitadas pelo sistema de billing.

Regras

Um dos momentos em que a fraude deve ser sinalizada é ao verificar o uso de um serviço a partir de uma linha telefônica que não possui habilitação para este tipo de serviço. Por exemplo, pode-se citar a ativação de serviços de Internet (WAP) diretamente nos switchs sem que o serviço possa ser faturado e cobrado.

3.5.4 A detecção de Fraude baseada em regras

Um dos objetivos desta análise é definir os perfis de fraudes em Telecomunicações através de regras, contendo uma ou mais condições, identificadas pela modelagem estatística. Assim, após a verificação de todas as condições sinaliza-se à empresa operadora a ocorrência de fraude em seu sistema por meio do disparo de um alarme (caracterizando-se por 1: chamada telefônica fraudulenta e 0: chamada telefônica não fraudulenta).

Os alarmes são acumulados em casos (um caso por conta telefônica) junto com os dados da conta telefônica e dos CDRs envolvidos na chamada telefônica fraudulenta. Estas informações servem de input para novas modelagens estatísticas, onde os modelos preditivos para propensão à fraude são constantemente atualizados e assim novos tipos de fraudes de sobreposição podem ser detectados.

Para que os modelos estatísticos sejam construídos é necessário que estes tenham como variáveis explicativas (ou input do modelo) informações detalhadas das chamadas; dos clientes e dos monitoramentos comportamentais destes.

O objetivo é que a modelagem estatística ofereça à empresa operadora informações mais completas que os tradicionais procedimentos de detecção de fraudes, já que analisa simultaneamente um maior volume de variáveis explicativas e permite análises de dados históricos dos tipos de fraudes e de seus alarmes.

O processo de descoberta de regras de fraudes pode ajudar a minimizar os falsos alarmes e auxilia a ajustar as regras existentes. Para isto, o processo deve observar as seguintes etapas:

- Um conjunto de regras identificado serve como base de dados para análise;
- Utiliza um dos procedimentos "*White Box*" ou "*Black Box*" para descobrir regras existentes no conjunto. O procedimento "*White Box*" tem fácil interpretação para associar as regras existentes, enquanto o "*Black Box*" tem uma interpretação mais complexa para o mesmo propósito.
- Disponibiliza associações entre as regras existentes, melhorando o sistema de análise de fraude. As descobertas são inseridas ao conjunto de regras a fim de melhorar cada vez mais o gerenciamento de fraudes em Telecomunicações.

Capítulo 4

Metodologia de Modelagem

As últimas décadas vêm mostrando um elevado aumento na quantidade de informações ou dados armazenados em formato eletrônico, consequência natural dos avanços tecnológicos e da crescente importância da informação no mundo real. Pela maneira em que os dados são guardados, e obviamente pelo grande volume, é impossível analisá-los e interpretá-los por métodos manuais, onde o especialista compara suas hipóteses com a massa de dados. Porém quanto maior for a quantidade de dados agrupados de forma lógica, maior a quantidade de informação armazenada, mesmo que codificada em símbolos e estruturas de dados aparentemente sem valor. Desta forma, surge a necessidade de se explorar estes dados para se extrair informações implícitas e utilizá-las no contexto do problema em questão.

O processo capaz de descobrir este conhecimento em banco de dados chama-se KDD (Knowledge Discovery in Database). é resultado da fusão de áreas como: Banco de Dados, Aprendizado de Máquina (uma sub-área de Inteligência Artificial), e é Estatística.

4.1 Descoberta de Conhecimento e Mineração de Dados

Durante os últimos anos tem se verificado um crescimento no volume de dados armazenados em meios magnéticos. Estes dados, produzidos e armazenados em larga escala, são inviáveis de serem lidos ou analisados por especialistas através de métodos ma-nuais

tradicionais[1], tais como planilhas de cálculos e relatórios informativos operacionais, onde o especialista testa sua hipótese contra a base de dados. Por outro lado, sabe-se que grandes quantidades de dados equivalem a um maior potencial de informação. Entretanto, as informações contidas nos dados não estão explicitamente caracterizadas. Desta forma, temos de explorar estes dados para extrair informação, isto é, o conhecimento implícito, e utilizá-la no âmbito do problema. Assim a necessidade de sistemas para dar suporte a decisão têm se desenvolvido com uma granularidade de informações mais refinada. Na década de 60 as exigências e necessidades estavam a nível de mercado; nos anos 70, ao nível de nichos, grupos de interesse; nos anos 80, a nível de segmentos de mercado; e nos anos 90, a nível de clientes. Este último nível, naturalmente, requer o uso de mais dados para se extrair conhecimento [2]. A exploração da informação neles contida implicitamente, depende de técnicas como Regras de Associação [3], Classificação [4], Clustering [5], entre outras, capazes de gerenciar tarefas complexas.

O processo capaz de descobrir este conhecimento em banco de dados chama-se KDD (*Knowledge Discovery Database*). O processo de KDD foi proposto em 1989 para referir-se às etapas que produzem conhecimentos a partir dos dados e, principalmente, à etapa de mineração dos dados, que é a fase que transforma dados em informações [6]. Este processo envolve encontrar e interpretar padrões nos dados, de modo iterativo e interativo, através da repetição dos algoritmos e da análise de seus resultados. Esse processo contém as seguintes fases:

- definição do problema;
- seleção dos dados;
- limpeza dos dados;
- pré-processamento dos dados;
- codificação dos dados;
- enriquecimento dos dados;
- mineração dos dados (Data Mining) e

- interpretação dos resultados.

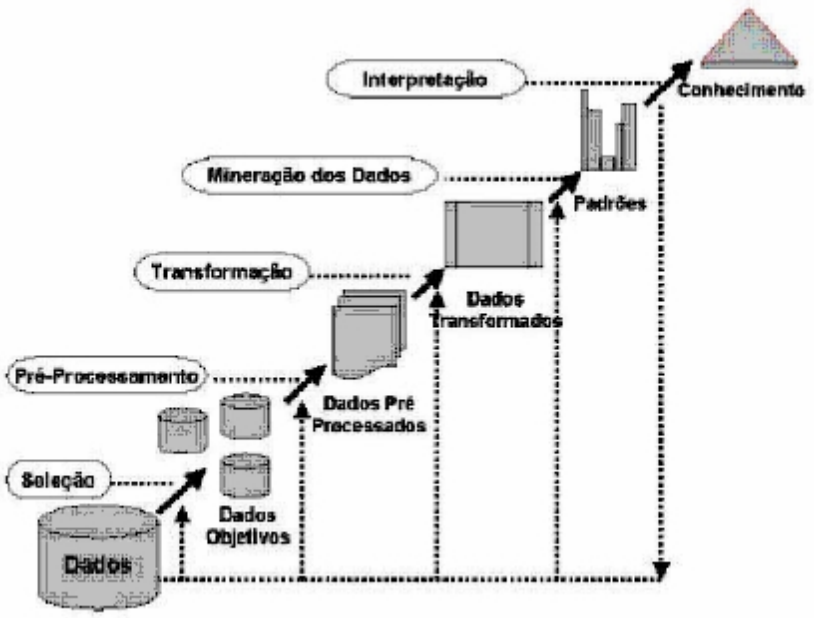


Figura 4.1: Processo KDD

O processo de KDD (figura 4.1) é formado pela intersecção de diferentes áreas. As áreas mais relacionadas em descoberta de conhecimento são: Machine Learning [7], [8], Inteligência Computacional, Estatística [9] e Visualização dos Dados [10]. Na área de Inteligência Computacional, em particular, as técnicas mais utilizadas são: Redes Neurais Artificiais [11], [12], Indução de regras [13] e Algoritmos Genéticos [14].

A Inteligência Computacional é uma área da ciência que busca, através de técnicas inspiradas na Natureza, o desenvolvimento de sistemas inteligentes que imitam aspectos do comportamento humano, tais como: aprendizado, percepção, raciocínio, evolução e adaptação.

Técnica	Inspiração
Redes Neurais	Neurônios Biológicos
Algoritmos Genéticos	Evolução Biológica
Lógica Fuzzy	Processo Linguístico
Sistemas Especialistas	Inferência

Redes Neurais são modelos computacionais não lineares, inspirados na estrutura e operação do cérebro humano, que procuram reproduzir características humanas, tais como: aprendizado, associação, generalização e abstração. Redes Neurais são ditas efetivas no aprendizado de padrões a partir de dados não lineares, incompletos, com ruído ou compostos de exemplos contraditórios.

Algoritmos Genéticos são algoritmos matemáticos inspirados nos mecanismos de evolução natural e recombinação genética. Esta técnica fornece um mecanismo de busca adaptativa baseada no princípio Darwiniano de reprodução e sobrevivência dos mais aptos.

Lógica Nebulosa (Fuzzy Logic) tem por objetivo modelar o modo aproximado de raciocínio humano, visando desenvolver sistemas computacionais capazes de tomar decisões racionais em um ambiente de incerteza e imprecisão. A Lógica Nebulosa oferece um mecanismo para manipular informações imprecisas, tais como os conceitos de muito, pouco, pequeno, alto, bom, quente, frio, etc, fornecendo uma resposta aproximada para uma questão baseada em um conhecimento que é inexato, incompleto ou não totalmente confiável.

Sistemas Especialistas são programas computacionais destinados a solucionar problemas em um campo especializado do conhecimento humano. Usa técnicas de Inteligência Artificial, base de conhecimento e raciocínio inferencial.

As técnicas da Inteligência Computacional têm sido empregadas no desenvolvimento de sistemas inteligentes de previsão, suporte à decisão, controle, otimização, modelagem, classificação e reconhecimento de padrões em geral, aplicados em diversos setores, tais como Energia, Industrial, Econômico, Financeiro, Comercial e Outros.

A mineração de dados é considerada a principal fase do processo de KDD. Essa fase é exclusivamente responsável pelo algoritmo minerador, ou seja, o algoritmo que diante da tarefa especificada, busca extrair o conhecimento implícito e potencialmente útil dos dados. A mineração de dados é, na verdade, uma descoberta eficiente de informações válidas e não óbvias de uma grande coleção de dados [27].

A proposta de extrair conhecimento de banco de dados surgiu devido a explosão do crescimento da quantidade de dados armazenados em meios magnéticos e da necessidade de aproveitá-los, motivada pela “fome de conhecimento”. Outro fator que contribuiu em muito para o aumento do interesse em mineração de dados foi o desenvolvimento das

técnicas de machine learning - redes neurais artificiais, algoritmos genéticos, entre outras, que tornaram a descoberta de relações interessantes em bases de dados mais atrativa.

Quando fala-se de mineração de dados não está se considerando apenas consultas complexas e elaboradas que visam ratificar uma hipótese gerada por um usuário em função dos relacionamentos existentes entre os dados, e sim da descoberta de novos fatos, regularidades, restrições, padrões e relacionamentos.

4.2 Redes Neurais

Uma Rede Neural Artificial (RNA) é uma técnica computacional que constrói um modelo matemático, emulado por computador, de um sistema neural biológico simplificado, com capacidade de aprendizado, generalização, associação e abstração. As RNAs tentam aprender padrões diretamente dos dados através de um processo de repetidas apresentações dos dados à rede, ou seja por experiência. Dessa forma, uma RNA procura por relacionamentos, constrói modelos automaticamente, e os corrige de modo a diminuir seu próprio erro.

Semelhante ao sistema biológico, uma RNA possui, simplificadamente, um sistema de neurônios, ou nós, e conexões ponderadas (equivalente às sinapses), pesos. Numa RNA os nós são arrumados em camadas, com conexões entre elas. A figura 4.2 representa conceitualmente a arquitetura de uma RNA simples. Os círculos representam os nós e as linhas representam os pesos das conexões. Por convenção, a camada que recebe os dados é chamada camada de entrada e a camada que mostra o resultado é chamada camada de saída. A camada interna, onde localiza-se o processamento interno, é tradicionalmente chamada de camada escondida. Uma RNA pode conter uma ou várias camadas escondidas, de acordo com a complexidade do problema [28],[29].

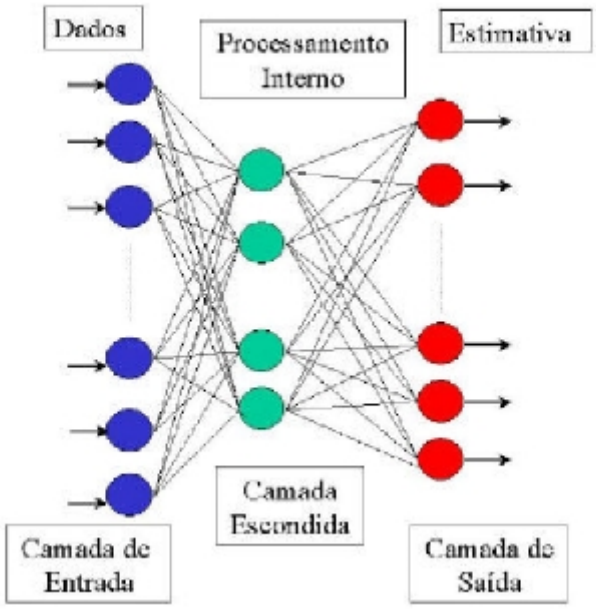


Figura 4.2: Arquitetura de uma Rede Neural Artificial simples

Para entender como uma RNA aprende é necessário saber como os pesos da rede afetam sua saída. O aprendizado de uma RNA envolve os ajustes dos pesos. A Figura 4.3 mostra o esquema de um neurônio artificial criado a partir do modelo simplificado do neurônio biológico [30]. O neurônio artificial possui várias entradas, que podem ser estímulos do sistema ou saídas de outros neurônios.

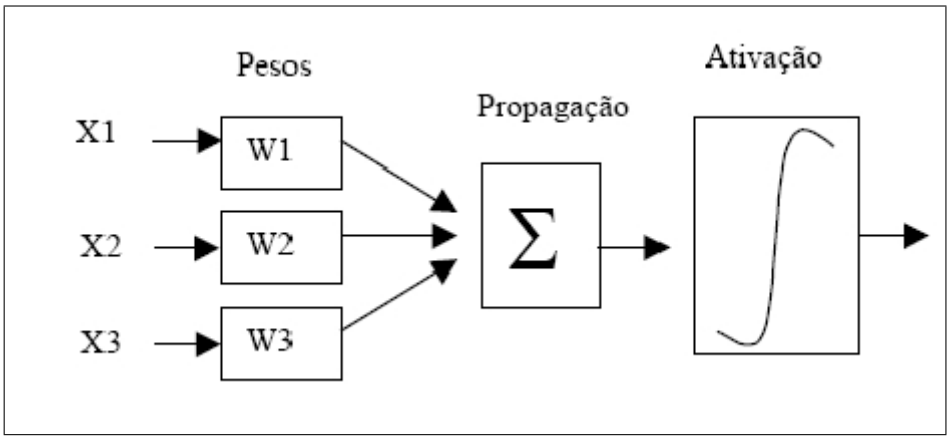


Figura 4.3: Esquema Simplificado de um neurônio artificial

O neurônio artificial é dividido em 2 seções funcionais. A primeira seção combina todas as entradas que alimenta o neurônio. Essa etapa indica como as entradas serão computadas (regra de propagação). A segunda seção recebe esse valor e faz um cálculo determinando o grau de importância da soma ponderada utilizando uma função de transferência, ou função de ativação. Essa função determina a que grau uma soma causará uma excitação ou inibição do neurônio. Os tipos mais comuns de funções de ativação são sigmóide e tangente hiperbólica, pois fornecem a característica de não linearidade para uma RNA.

Uma RNA ajusta seus pesos na fase de treinamento. É fornecido um dado de observação, o qual é processado, e uma resposta será produzida. O resultado fornecido é comparado com uma saída desejada, saída correta. Se a rede acerta essa saída, então ela não faz nada, entretanto se o resultado não está correto, ocorre um ajuste dos pesos de modo que o erro seja minimizado.

As topologias mais comuns de RNAs são as de múltiplas camadas feed-forward e as redes recorrentes. O aprendizado de uma RNA pode ser dividido em 3 grupos: sem treinamento – os valores dos pesos sinápticos são estabelecidos a priori, ou seja, ajustados em um único passo, por exemplo, Redes de Hopfield [31]; treinamento supervisionado – a rede é treinada através do fornecimento dos valores de entrada e dos seus respectivos valores de saída desejados (procura minimizar o erro médio quadrado); e treinamento nãosupervisionado – o sistema extrai as características dos dados fornecidos, agrupando-os em classes (clusters).

As principais aplicações de Redes Neurais em mineração são classificação, clustering, aproximação de funções, previsão e verificação de tendências. De um modo geral, a arquitetura de uma RNA recebe uma tupla (atributos preditivos, ou seja, atributos que pertencem a parte SE de uma regra) como entrada através da primeira camada da rede (camada de entrada).

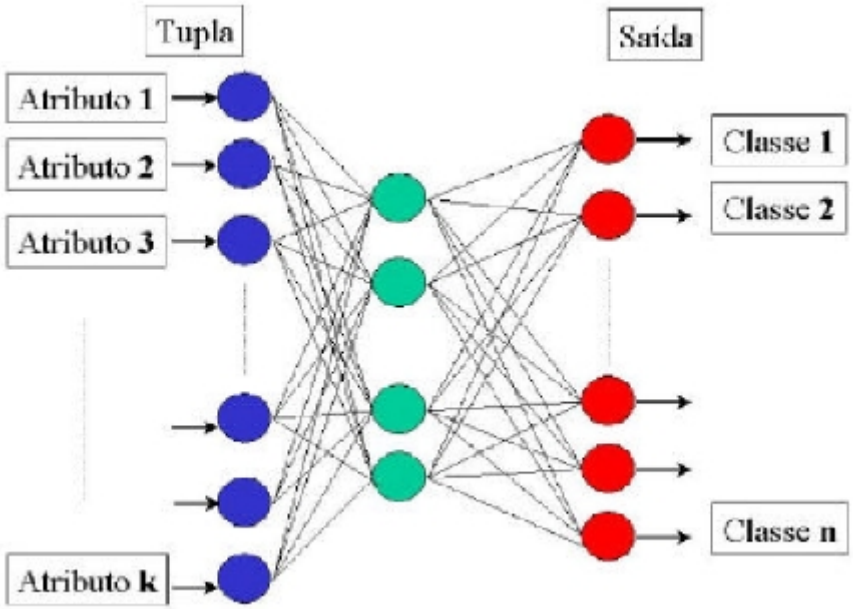


Figura 4.4: Modelo de um Rede Neural Artificial para mineração de dados

Tratando-se de aprendizado não supervisionado, não existe atributos objetivo que possa ser utilizado para corrigir os pesos da rede. Esse tipo de aprendizado aplica-se em tarefas de clustering. As redes auto organizáveis (por exemplo, Kohonen), baseadas em aprendizado competitivo, destacam-se como um bom algoritmo [32].

Entretanto, em algoritmos supervisionados, os atributos objetivo são modelados pela camada de saída da rede. Deste modo o algoritmo pode estimar o quanto a saída desejada está distante da saída real. O algoritmo mais comum em RNAs com aprendizado supervisionado é o back progation. Seu objetivo é minimizar a função erro entre a saída real da rede e a saída desejada utilizando o método do gradiente descendente [33].

Back-propagation é utilizado para classificar, aproximar funções, prever e verificar tendências [34]. Tomando como exemplo a Figura 4.4 , a camada de entrada pode ser tal que cada nó representa um determinado atributo preditivo de uma tupla e a camada de saída decodifica a que classe essa tupla pertence, ativando um único nó. Maiores detalhes podem ser encontrados em [35]. Para a tarefa de classificação também são utilizadas as redes neurais probabilísticas, baseadas em classificadores bayesianos e as redes RBF

(Radio-Basis Function), baseadas em funções gaussianas. Esses algoritmos geram curvas de densidade de probabilidade, fornecendo resultados com bases estatísticas. Esses resultados indicam o grau de evidência sobre o qual se baseia a decisão. Entretanto, essa metodologia só funciona bem se existir um número suficiente de exemplos na base de dados. A Figura 4.5 se enquadra como arquitetura desses algoritmos. A principal diferença está na interpretação do resultado, pois cada nó da camada de saída gera um valor que indica a probabilidade da tupla inserida na camada de entrada pertencer a uma determinada classe.

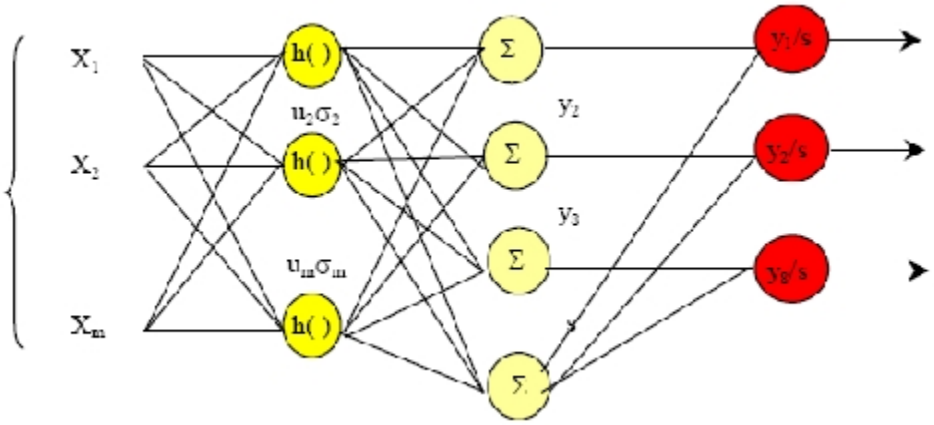


Figura 4.5: Topologia de uma RNA utilizando bases estatísticas

Nesta topologia, RBF, o número de neurônios da camada escondida é exatamente igual ao número de tuplas apresentadas para treinamento. Onde: X_i é um atributo preditivo da tupla apresentada, $h()$ é a função de ativação, u_i é o centro da Gaussiana e σ_i é o desvio padrão. As redes Hopfield também são utilizadas para classificação. Nesse algoritmo as classes são consideradas estados. Fornecendo-se uma tupla como entrada para a rede, os pesos são atualizados de modo a ocorrer uma convergência para um estado estável, que será sua classe. Tresp destaca o grande interesse em entender o conhecimento intrínseco que a rede neural adquire no treinamento [36]. Nesse trabalho argumenta-se que redes com funções de base gaussiana podem ser geradas de simples regras probabilísticas e, também, se regras apropriadas de aprendizados são utilizadas, regras probabilísticas podem ser extraídas das redes treinadas.

Towell apresenta um método que eficientemente extrai regras simbólicas de uma RNA

treinada [37]. Os resultados obtidos através de testes empíricos desse método permitem concluir que as regras extraídas:

1.
   - reproduzem com proximidade a acurácia da rede a qual foi extraída;
   - são superiores às regras produzidas por métodos que diretamente refinam regras simbólicas;
   - são superiores àquelas produzidas por técnicas anteriores para extração de regras de RNAs treinadas; e
   - são compreensíveis.

Conclui-se que esse método demonstra que uma RNA pode ser utilizada para efetivamente refinar o conhecimento simbólico.

4.3 Árvore de Decisão

As árvores de decisão/classificação são técnicas utilizadas na construção de modelos de análise de dados e na sua classificação. Uma das principais características de uma árvore de decisão é seu tipo de representação: estrutura hierárquica que traduz uma árvore invertida e que se desenvolve da raiz para as folhas. A representação hierárquica traduz uma progressão da análise de dados no sentido de desempenhar uma tarefa de previsão/classificação. Em cada nível da árvore tomam-se decisões acerca da estrutura do nível seguinte até atingir os nós terminais (nós folhas).

O princípio inerente desta modelagem é a divisão/classificação do evento analisado, de forma que em cada nível da árvore a tarefa de predizer/classificar uma massa com alta dispersão da varíavel resposta é decomposta em subproblemas mais homogêneos. Esse conceito é traduzido na estrutura da árvore, à medida que aumentam os nós atenua-se a heterogeneidade da variável resposta, onde as previsões podem ser feitas com risco menor de erro. É uma modelagem que se desenvolve do geral para o específico, onde cada novo nível descendente particulariza o valor das variáveis explicativas.

A Árvore de Decisão é definida, então, como uma estrutura de dados recursivamente definida com nós folha, que indicam uma classe, ou nós de decisão que contem um teste sobre o valor do atributo. Para cada um dos possíveis valores de um atributo tem-se um

ramo para outra árvore de decisão, que contém a mesma estrutura da árvore principal. Este tipo de aboradgem divide o espaço de descrição do problema conjuntos disjuntos. ë um método de classificação supervisionado, onde a varíavel resposta é explicada a partir de n variáveis independentes.

As árvores de decissão/classificação podem ser usadas com objetivos distintos, de acordo com o problema que se quer analisar. Pode-se classificar as informações de uma população, descobrir a estrutura de determinado problema, identificar as variáveis que interferem na sua resolução e construir um modelo que o solucione. Neste tipo de metodologia é possível identificar as variáveis explicativas mais relevantes para descrever determinada situação.

As principais vantagens na utilização desta metodologia são:

- Ausência de pressupostos inerentes aos modelos paramétricos em problemas com elevado número de variáveis explicativas;
- Árvores de decisão ou de classificação são técnicas de indução usadas para descobrir regras de classificação para um atributo a partir da subdivisão sistemática dos dados contidos no repositório que está sendo analisado.
- As árvores de decisão consistem de nodos que representam os atributos, de arcos, provenientes destes nodos e que recebem os valores possíveis para estes atributos, e de nodos folha, que representam as diferentes classes de um conjunto de treinamento.

Uma árvore de decisão tem a função de particionar recursivamente um conjunto de treinamento, até que cada subconjunto obtido deste particionamento contenha casos de uma única classe. Para atingir esta meta, a técnica de árvores de decisão examina e compara a distribuição de classes durante a construção da árvore. O resultado obtido, após a construção de uma árvore de decisão, são dados organizados de maneira compacta, que são utilizados para classificar novos casos [39].

A partir de uma árvore de decisão é possível derivar regras. As regras são escritas considerando o trajeto do nodo raiz até uma folha da árvore. As regras e a árvore de decisão são geralmente utilizadas em conjunto. Devido ao fato das árvores de decisão tenderem a crescer muito, de acordo com algumas aplicações, elas são muitas vezes substituídas pelas

regras Isto acontece em virtude das regras poderem ser facilmente modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de se referenciar outras regras [40].

A utilização de árvores de decisão binárias para classificação pode ser considerada uma abordagem não-paramétrica para reconhecimento de padrões. Uma árvore de decisão faz uma representação hierárquica do espaço de feições, onde padrões x_i são alocados às classes $w_j (j = 1, 2, ..., k)$ conforme o resultado encontrado depois de percorridos os ramos da árvore. Estevtipo de árvore de decisão foi discutido em profundidade por Breiman [41], cujas contribuições são associadas à sigla CART (classification and regression trees).

Árvores de decisão binárias consistem de repetidas divisões do espaço de feições em dois subgrupos descendentes que terminam em nodos associados às classes w_j. Na terminologia de árvores de decisão os subgrupos do espaço de feições são definidos através de nodos. Uma árvore de decisão com alto poder preditivo e um pequeno número de nodos constitui uma situação altamente desejável.

A técnica CART constrói uma árvore de decisão binária a partir de uma amostra de treinamento formando partições na amostra. Esse fato faz com que os tamanhos amostrais dos subgrupos formados por tais partições decresçam, exigindo um tamanho amostral razoável para obtenção de bons resultados, conforme sugere McLachlan[42].

A árvore de decisão inicia com o nodo raiz t contemplando aquela variável do espaço de feições que minimiza o grau de impureza de dois nodos irmãos. De acordo com Breiman [41] a medida de impureza do nodo t – denotada por $i(t)$– pode ser determinada por:

$$i(t) = -\sum_{j=1}^{k} p(w_j|t) \log p(w_j|t) \tag{4.1}$$

onde $p(w_j|t)$ é a proporção de padrões xi alocados à classe wj no nodo t. Todos nodos não terminais dividem-se em outros dois nodos, digamos t_L e t_R, sendo que p_L é a proporção de observações que seguem para o nodo t_L e p_R a proporção de casos que segue para t_R. A melhor divisão s é aquela que torna máxima a diferença:

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{4.2}$$

O crescimento da árvore de decisão se dá por sucessivas divisões até que não haja mais

possibilidade de decréscimo significante no grau de impureza por meio de uma divisão s. Quando isto ocorre, o nodo t não é mais dividido, sendo automaticamente transformado em um nodo terminal. A classe w_j associada ao nodo terminal t é aquela que maximiza a probabilidade condicional $p(w_j|t)$.

Segundo Brazdil [38] muitos são os algoritmos de classificação que elaboram árvores de decisão. Não há uma forma de determinar qual é o melhor algoritmo, um pode ter melhor desempenho em determinada situação e outro pode ser mais eficiente em outros tipos de situações. O algoritmo ID3 foi um dos primeiros algoritmos de árvore de decisão, tendo sua elaboração baseada em sistemas de inferência e em conceitos de sistemas de aprendizagem. Logo após foram elaborados diversos algoritmos, sendo os mais conhecidos: C4.5, CART (Classification and Regression Trees), CHAID (ChiSquare Automatic Interaction Detection), entre outros. Os algoritmos que constroem árvores de decisão buscam encontrar aqueles atributos e valores que provêem máxima segregação dos registros de dados, com respeito ao atributo que se quer classificar, a cada nível da árvore.

4.4 Regressão Logística

Em qualquer problema de regressão a quantidade de interesse é o valor médio da variável resposta, dado o valor da variável independente. Esta quantidade é normalmente designada por média condicional e pode ser expressa por $E(Y|x)$, onde Y designa a variável resposta e x designa o valor da variável independente.

Numa regressão linear assume-se que esta média pode ser expressa como uma equação linear em x, do tipo:

$$E(Y|x) = \beta_0 + \beta_1 x \tag{4.3}$$

A partir desta expressão, verifica-se que $E(Y|x)$ pode tomar qualquer valor com x a variar de $-\infty$ a $+\infty$.

Considere-se $\pi(x) = E(Y|x)$. A forma específica do modelo de regressão logística para uma variável resposta dicotômica, tem a forma:

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} \tag{4.4}$$

A transformação de $\pi(x)$ é denominada por transformação logit. Esta transformação é definida em termos de $\pi(x)$, como sendo:

$$g(x) = \ln\left(\frac{\pi(x)}{1 + \pi(x)}\right) = \beta_0 + \beta_1 x \tag{4.5}$$

A importância desta transformação é que $g(x)$ tem muitas propriedades desejáveis dos modelos de regressão linear. O logit $g(x)$ é linear nos seus parâmetros, é uma função contínua, e pode variar de $-\infty$ a $+\infty$, consoante o domínio de variação de x.

Uma diferença importante entre os modelos de regressão linear e o da regressão logística diz respeito à distribuição condicional da variável resposta.

Na regressão linear assume-se que uma observação da variável resposta pode ser expressa como $y = E(Y|x) + \varepsilon$, em que ε é designado por erro, e dá o desvio de uma observação em relação à média condicional. A hipótese mais comum é que este erro ε segue uma distribuição Normal com média zero e variância constante ao longo dos níveis da variável independente; assim resulta que a distribuição da variável resposta dado x, será Normal com média $E(Y|x)$, e variância constante. Quando a variável resposta é dicotômica, este pressuposto não se verifica. Nesta situação, deve-se expressar o valor da variável resposta dado x como $y = \pi(x) + \varepsilon$. Aqui a quantidade ε pode assumir um dos dois valores possíveis:

$$\begin{aligned} Y = 1 &\Longrightarrow \varepsilon = 1 - \pi(x) \quad \text{com probabilidade } \pi(x) \\ Y = 0 &\Longrightarrow \varepsilon = -\pi(x) \quad \text{com probabilidade } 1 - \pi(x) \end{aligned} \tag{4.6}$$

Então, ε tem uma distribuição com média zero e variância igual a $\pi(x)[1 - \pi(x)]$, isto é, a variável resposta segue uma distribuição binomial com probabilidade dada pela média condicional, $\pi(x)$.

4.4.1 Ajuste do modelo de Regressão Logística

Considere uma amostra de n observações independentes do par (x_i, y_i) com $i = 1, 2, \cdots, n$, e yi e xi, designam, respectivamente, o valor da variável resposta e o valor da variável

independente, correspondente ao i-ésimo indivíduo.

No ajuste um modelo de regressão logística estimam-se os parâmetros desconhecidos, β_0 e β_1. Na regressão linear o método mais utilizado é o dos mínimos quadrados, onde estimam-se os valores de β_0 e β_1 que minimizam a soma dos quadrados dos desvios dos valores observados de Y em relação aos valores previstos baseados no modelo especificado. Sobre as usuais condições para a regressão linear, o método dos mínimos quadrados conduz a estimadores com um número de propriedades estatísticas desejáveis. Infelizmente, quando este método é aplicado a um modelo de resposta dicotômica, os estimadores não apresentam as mesmas propriedades, pois o pressusposto de erros normalmente distribuídos não pode ser verificado. A alternativa é utilizar o método da máxima verossimilhança. Para aplicar tal metodologia é necessário construir a função de verossimilhança.

A função de verossimilhança, expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores de máxima verossimilhança (EMV) destes parâmetros, são escolhidos de forma a maximizarem a função de verossimilhança.

Para o modelo de regressão logística dicotômica, onde a variável resposta é codificada por $Y = 0$ e $Y = 1$, a função de probabilidade condicional pode ser expressa através de:

$$P(Y/x) = \begin{cases} \pi(x) & \text{se} \quad Y = 1 \\ 1 - \pi(x) & \text{se} \quad Y = 0 \end{cases} \tag{4.7}$$

Assim, para os pares (x_i, y_i), quando $y_i = 1$ a contribuição para a função de verossimilhança é $\pi(x_i)$, e para os pares cujo valor $y_i = 0$ a contribuição para a função de verossimilhança é $1-\pi(x_i)$, onde a quantidade $\pi(x_i)$ designa o valor de $\pi(x)$ calculada num valor x_i. Uma forma de expressar a contribuição para a função de verossimilhança do par (x_i, y_i) é através do termo:

$$\zeta(x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{4.8}$$

Desde que as observações sejam independentes, a função de verossimilhança é obtida por:

$$l(\beta) = \prod_{i=1}^{n} \zeta(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{4.9}$$

onde $\pi(x_i)$ representa $P(Y = 1|xi)$, também designada por probabilidade de sucesso. O método da máxima verossimilhança utiliza os valores estimados de β que maximize a expressão (4.14) . Matematicamente, torna-se mais fácil trabalhar a expressão do logarítmo da verossimilhança, dada por

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \cdot ln[\pi(x_i)] + (1 - y_i) \cdot ln[1 - \pi(x_i)]\} \tag{4.10}$$

Para achar o valor de β que maximiza $L(\beta)$, deriva-se a expressão acima em ordem a cada parâmetro e se igualam as expressões obtidas a zero, obtendo-se assim as equações de verossimilhança:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] \tag{4.11}$$

e

$$\frac{\partial L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} x_i \cdot [y_i - \pi(x_i)] \tag{4.12}$$

Para a regressão logística dicotômica, as equações de verossimilhança são não lineares em β , o que requer métodos de resolução de equações não lineares do tipo Newton-Raphson.

4.4.2 Descrição da natureza do evento de fraude

A fraude de subscrição é uma variável binária Y_i que assume o valores 0 ou 1. Consideramos que Y_i é uma variável aleatória de Bernoulli, cuja distribuição de probabilidade é:

Y_i	Probabilidade
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Onde π_i é a probabilidade de que o cliente seja fraudador, e $1 - \pi_i$ a probabilidade de que ele não seja fraudador. Neste caso a forma da função resposta é frequentemente

sigmoidal (figura 4.6).

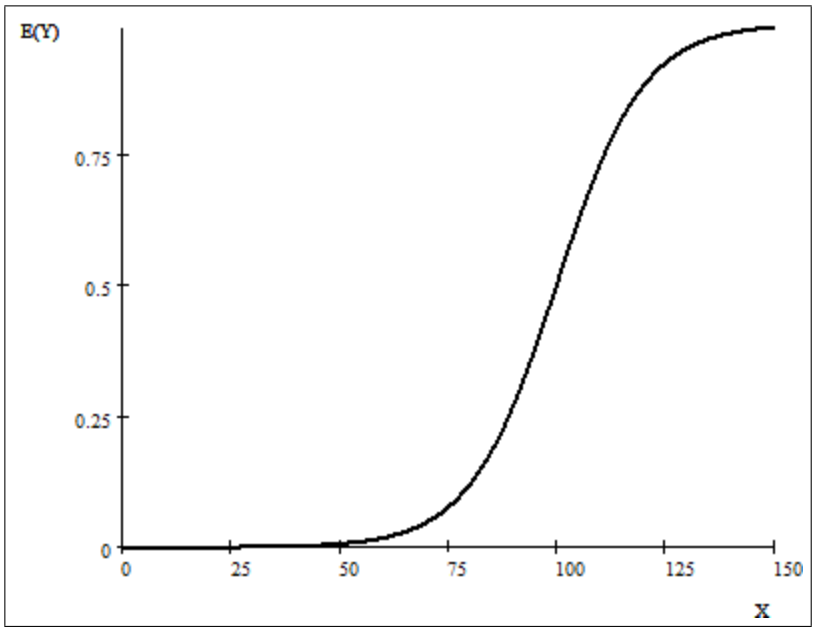


Figura 4.6: Exemplo de função logística resposta

A fraude de subscrição para pré-ativação considera um conjunto de até 39 covariadas.

No caso específico em estudo, a modelagem identificou 14 variáveis significativas para determinar a propensão à fraude de subscrição. Com a intenção de simplificar utilizaremos a notação matricial:

$$\underset{px1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{px1}{X} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix} \tag{4.13}$$

onde $\beta' X = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$e a função logística passa a ser:

$$E(Y) = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)} \tag{4.14}$$

considerando a função de ligação logito:

$$\pi' = \log\left(\frac{\pi}{1-\pi}\right) \tag{4.15}$$

Capítulo 5

Curva ROC (Receiver Operating Characteristic)

A análise ROC (Receiver Operating Characteristic) é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico em medicina. Esta análise por meio de um método gráfico simples e robusto, permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte. Neste capítulo fazemos a descrição desta metodologia de análise.

5.1 Introdução

A análise ROC (Receiver Operating Characteristic) teve origem na teoria de decisão estatística e foi desenvolvida entre 1950 e 1960 para avaliar a detecção de sinais em radar e na psicologia sensorial . A potencial utilidade da análise ROC em avaliar diagnósticos médicos foi desde então utilizada por vários autores [19] e, subsequentemente, foi aplicada com sucesso a uma grande variedade de testes de diagnóstico [18], e em particular no diagnóstico de imagem médica.

Charles E. Metz desenvolveu um conjunto de trabalhos sobre a aplicabilidade da análise ROC a sistemas de diagnóstico, nomeadamente no campo da imagem radiológica. Em [21] apresenta alguns princípios básicos da análise ROC, como o significado de sensibilidade e especificidade no desempenho dos testes e diagnóstico.

Define sensibilidade e especificidade como duas medidas de precisão de um teste de

diagnóstico, dadas pelas frações:

$$\text{SENSIBILIDADE} = \frac{\text{nº de decisões verdadeiras positivas}}{\text{nº de casos realmente positivos}}$$

e

$$\text{ESPECIFICIDADE} = \frac{\text{nº de decisões verdadeiras negativas}}{\text{nº de casos realmente negativos}}$$

5.2 Teoria de análise ROC

Define também, valor de corte, como sendo um valor que pode ser selecionado arbitrariamente entre os valores possíveis para a variável de decisão, e acima do qual classifica como positivo (teste de diagnóstico positivo, presença de doença), e abaixo do qual classifica como negativo (teste de diagnóstico negativo, ausência de doença).

Assim, se existir alguma sobreposição entre a distribuição dos casos classificados como positivos e a distribuição dos casos classificados como negativos, e forçando o valor de corte a percorrer todos os valores possíveis da variável de decisão, podem-se obter vários pares de frações de verdadeiros positivos (sensibilidade) e de falsos positivos (1 - especificidade), que corresponderão, segundo Metz [21], aos eixos coordenados "y" e "x" de um gráfico que este designou por curva ROC para o teste de diagnóstico. Esta curva pode descrever as características de detecção associadas ao teste, e o observador pode operar em qualquer ponto da curva desde que selecione o valor de corte apropriado de decisão.

Para Metz uma curva ROC convencional descreve os compromissos que podem ser tomados entre a FVP e a FFP, com a variação dos diferentes valores de corte ou critérios de decisão. Metz afirma que a análise ROC fornece uma descrição da capacidade de detectar o evento analisado independentemente da prevalência e dos efeitos de escolha do critério de decisão.

O problema em termos de testes de hipóteses, ou tomada de decisões estatísticas, é representado pela figura 5.1. A distribuição da esquerda representa nesta situação, a hipótese nula, H_0, e a da direita uma hipótese alternativa, H_1.

Assim, as hipóteses do problema poderão ser especificadas como:

H_0: A população tem média $\mu = \mu_0$;

H_1: A população tem média $\mu = \mu_1$.

Com base numa observação x, uma das hipóteses é aceite.

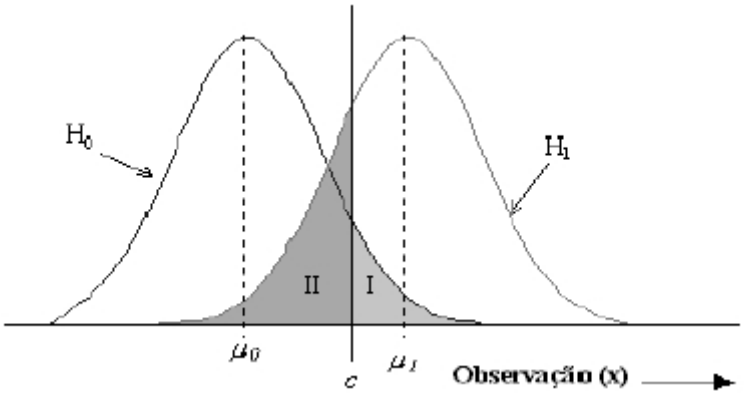


Figura 5.1: Distribuições de duas populações

Como pode ser verificado na figura 5.1, H_0 é a hipótese nula que considera que a população tem média $\mu = \mu_0$ e H_1 é a hipótese alternativa que considera que a população tem média $\mu = \mu_1$. Assim, a área sombreada à direita do critério de decisão, c, representa a probabilidade de cometer um *erro tipo I*, que corresponde à probabilidade de rejeitar H_0 quando H_0 é verdadeira; a área sombreada à esquerda do critério de decisão, c, representa a probabilidade de cometer um *erro tipo II*, que corresponde à probabilidade de não rejeitar H_0 quando H_1 é verdadeira.

A construção do teste estatístico é equivalente a dividir o eixo x em duas regiões, separadas pelo critério de decisão c. Valores de x menores que conduzirão à aceitação da hipótese nula, H_0, e valores de x maiores que conduzirão à aceitação da hipótese alternativa, H_1. Consoante o critério de decisão escolhido, pode-se determinar a probabilidade de cometer um *erro tipo I* ou *tipo II* (figura 5.1).

Existem princípios gerais para os testes de hipóteses que obedecem a determinadas regras desenvolvidas por Neyman e Pearson. A principal regra associada a estes, e a mais familiar em estatística, é fixar a probabilidade de cometer um *erro tipo I* arbitrariamente (a um nível de significância usualmente de 0.05 ou 0.01) e depois escolher um critério de forma a minimizar a probabilidade de cometer um *erro tipo II*. Estes autores demonstraram que o melhor teste é definido em termos da razão da verossimilhança. Aceita-se

H_1 quando a razão das verossimilhanças excede determinado valor c, que é escolhido para produzir a probabilidade desejada de cometer um *erro tipo I*.

A potência do teste é definida por:

$$k = \begin{cases} Prob\ (erro\ \ tipo\ I)\ \ sob\ H_0 \\ 1 - Prob(erro\ \ tipo\ II)\ \ sob\ H_1 \end{cases}$$

Sob as regras de Neyman-Pearson [23], fixa-se a probabilidade de cometer um *erro tipo I* e escolhe-se a razão de verossimilhança igual a c de forma a maximizar a potência do teste. Assim é possível definir a curva característica de operação, que não é mais do que a representação gráfica do complementar da função potência do teste $(1 - k)$. A curva ROC é uma maneira gráfica de comparar duas curvas características de operação - a que se definiu anteriormente, em que se fixa a probabilidade de cometer um *erro tipo I* arbitrariamente, e uma outra que mostra a variação em probabilidade de um *erro tipo I* para um valor fixo de probabilidade de cometer um *erro tipo II*. Na figura 5.2, encontram-se representadas as curvas características de operação para as duas situações descritas, considerando um teste hipotético para duas distribuições Normais com igual variância.

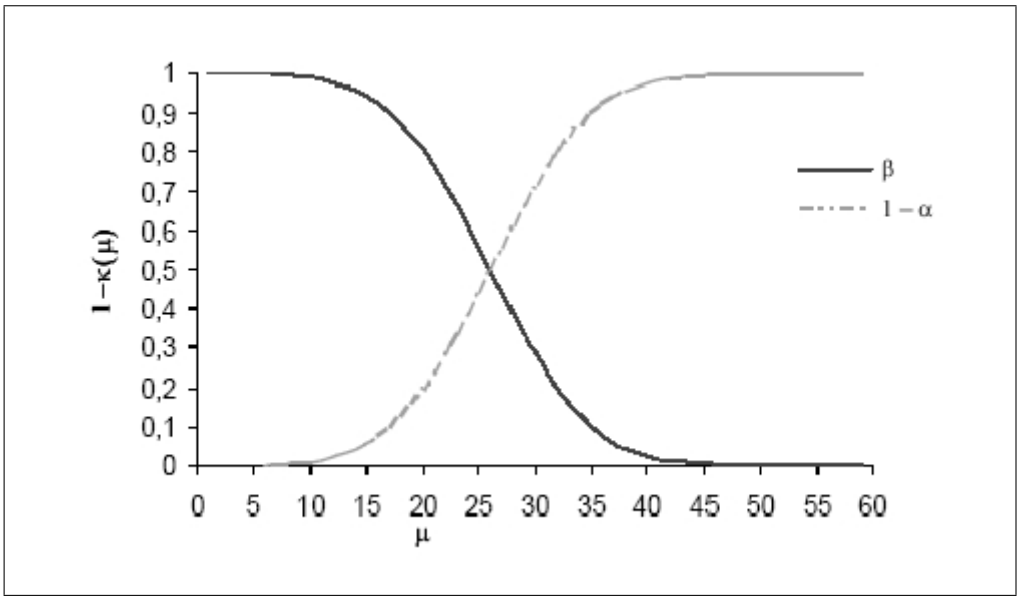


Figura 5.2: Representação de curvas características de operação

A curva ROC, transmite a informação da conjunção destas duas curvas de operação, isto é, mostra como podem variar os dois tipos de erro, com a mudança de critério de decisão. Um outro avanço na teoria da decisão estatística, por volta de 1940, foi dado por Abraham Wald. Wald demonstrou que algumas regras de decisão diferentes - como a maximização da proporção de decisões corretas, maximização do valor esperado de uma decisão e maximização da mínima recompensa são unificadas pela razão das verossimilhanças [18]. Posteriormente, Green e Sweets , descrevem algumas das regras de decisão mais utilizadas em estatística que serão apresentadas em seguida.

5.2.1 Maximização de uma combinação ponderada.

No caso da existência de duas alternativas, os resultados poderão ser descritos por quatro probabilidades diferentes. Apenas duas dessas probabilidades são independentes, dado que:

$$P(H_0|h_0) + P(H_1|h_0) = 1$$

e

$$P(H_0|h_1) + P(H_1|h_1) = 1$$

Assim, o objetivo seria, sempre que possível, maximizar $P(H_1|h_1)$, ao mesmo tempo que se minimizaria $P(H_1|h_0)$. Geralmente, não se consegue satisfazer os dois objetivos simultaneamente, pelo que se opta pela maximização da quantidade:

$$\{P(H_1|h_1) - \gamma P(H_1|h_0)\}$$

onde γ é uma constante, $\gamma > 0$.

Designando por A o conjunto de todos acontecimentos que conduzem à aceitação de h_1, então a probabilidade de H_1 ser aceita quando h_1 é verdadeira é dada por,

$$\sum_{ei \in A} P(e_i|h1) = P(H_1|h_1) \quad \text{(para o caso discreto)}$$
$$\int_{ei \in A} P(e_i|h_1) = P(H_1|h_1) \quad \text{(para o caso contínuo).}$$

De forma análoga, a probabilidade de uma aceitação incorreta da hipótese h1, é dada por,

$$\sum_{ei \in A} P(e_i|h_0) = P(H_1|h_0) \quad \text{(para o caso discreto)}$$
$$\int_{ei \in A} P(e_i|h_0) = P(H_1|h_0) \quad \text{(para o caso contínuo).}$$

Deve-se escolher a região A de forma a maximizar $P(H1|h1)$, :

$$\begin{aligned} P(H1|h1) - \gamma P(H1|h0) &= \sum_{ei \in A} P(e_i|h_1) - \sum_{ei \in A} P(e_i|h_0) \quad \text{para o caso discreto} \\ &= \int_{ei \in A} P(e_i|h_1) - \int_{ei \in A} P(e_i|h_0) \quad \text{para o caso contínuo} \end{aligned}$$

Note-se que apenas se deve incluir em A, acontecimentos cuja razão de verossimilhanças de um acontecimento e_k para a hipótese h_1 em relação à hipótese h_0 - $ln(e_k)$- satisfaçam a condição:

$$ln(e_k) = \frac{P(e_k|h_1)}{P(e_k|h_1)} \geq \gamma.$$

Assim, a primeira regra de decisão pode ser definida da seguinte forma:

Uma regra de decisão que maximize $P(H_1|h_1) - \gamma P(H_1|h_0)$, consiste em escolher H_1 se e só se a razão de verossimilhanças para todos acontecimentos e_i, $ln(e_i) \geq \gamma$, onde γ é o valor do critério adotado.

5.2.2 Maximização da porcentagem de respostas corretas.

Considerando que os custos associados aos erros são nulos e o valor de uma decisão correta igual a um, maximizar o valor esperado de uma estratégia de decisão é equivalente a maximizar a percentagem de respostas corretas.

Atendendo estas condições,

$$\gamma = \frac{P(h0)}{P(h1)}$$

se $P(h1)$ aumentar, é necessário uma menor razão de verossimilhanças para que H_1 seja escolhido.

5.3 Análise ROC

A detecção de sinais eletromagnéticos na presença de um ruído foi analisada, em 1940 como um problema de teste de hipóteses estatísticas. O ruído foi identificado como sendo a hipótese nula, H_0, enquanto o ruído mais sinal estava associado com a hipótese alternativa, H_1.

Por exemplo, no contexto dos radares, os erros de *tipo I* são designados "falsos alarmes", enquanto que os erros de *tipo II* são "falhas", e ambos são considerados perigosos numa situação de defesa, dado que os seus custos variam com os diferentes tratamentos e as reações disponíveis ao tratamento.

Na teoria de detecção do sinal, o observador tem como tarefa, decidir com base na aleatoriedade, qual dos estímulos é resultado do ruído mais sinal, ou do ruído. O problema fundamental de detecção, pode ser visto da seguinte forma [24]:

- Existe uma ocorrência aleatória de dois acontecimentos, ruído mais sinal (sn) e ruído (n), e cada acontecimento ocorre num intervalo de tempo bem definido;
- O estímulo físico, ou evidência relativo a cada acontecimento, varia de experiência para experiência, e tem um resultado, que é a representação probabilística do acontecimento;
- Após cada observação, o observador deve tomar uma decisão do tipo "sim" ou "não".

Assim, o procedimento de decisão, envolve dois elementos básicos: acontecimento $\rightarrow$ decisão. Cada estímulo deve ser classificado em uma de duas categorias, sn ou n.

Designando por $P(sn)$, a probabilidade associada à presença de sinal, e $P(n)$, a probabilidade associada à ausência de sinal (só ruído), no caso de dois acontecimentos:

$$P(sn) + P(n) = 1$$

Estas probabilidades são usualmente dadas pela experiência ou natureza e, normalmente, não se encontram sob controle do observador.

Um modelo do tipo acontecimento $\rightarrow$ decisão, poderá ser descrito em termos de árvore de probabilidades como ilustra a figura 5.3.

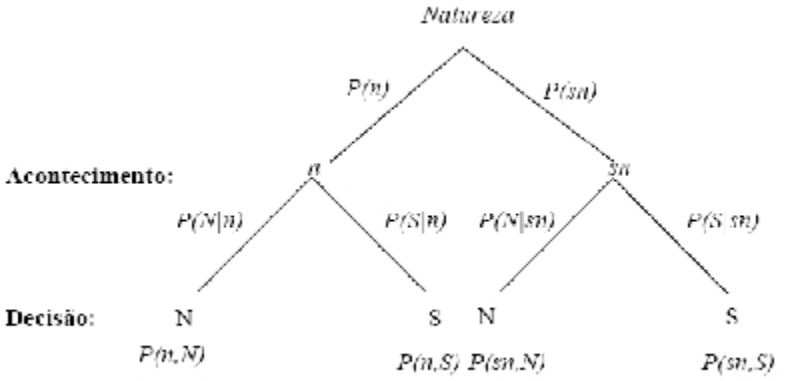


Figura 5.3 Árvore de probabilidades para o procedimento de detecção de sinal

Nesta situação a decisão do observador é do tipo: "sim, o sinal encontra-se presente", S ou "não, o sinal encontra-se ausente", N.

O desempenho de um observador numa experiência num único intervalo é usualmente medido em termos de probabilidades conjuntas de acontecimento-resposta. Estas probabilidades são baseadas quer no valor da probabilidade a priori da existência de sinal, $P(sn)$, quer nos valores das duas probabilidades condicionadas, $P(S|sn)$ e $P(S|n)$. Assim, define-se:

- aceitação correta: $P(sn, S) = P(S|sn)P(sn)$;
- rejeição incorreta: $P(sn, N) = [1 - P(S|sn)]P(sn)$;
- rejeição correta: $P(n, N) = [1 - P(S|n)]P(n)$;
- aceitação incorreta: $P(n, S) = P(S|n)P(n)$.

Assim, a ROC provêm de uma tabela de contingência $2 \times 2, dotipo$:

		Autoconhecimento	
		Ruído+Sinal (sn)	Ruído (n)
Decisão	Sim	$a = P(S/sn)$	$b = P(S/n)$
		Verdadeiro Positivo	Falso positivo
	Não	$c = P(N/sn)$	$d = P(N/n)$
		Falso negativo	Verdadeiro Negativo

A ROC é assim baseada em duas quantidades que contêm toda a informação da tabela de contingência, uma designada por fração de verdadeiros positivos (FVP), definida por $a/(a + c)$, e outra designada por fração de falsos positivos (FFP), definida por $b/(b + d)$, a fração de falsos negativos e a fração de verdadeiros negativos são os respectivos complementares.

Pode-se definir a ROC (Receiver Operating Characteristic) de duas formas diferentes, uma mais restritiva, em termos da razão de verossimilhanças, e uma outra mais geral, em termos da variável de decisão x [24].

Definição de ROC em termos de $l(x)$ - Uma ROC sumariza oconjunto possível de $matrizes2\times2$, que resulta quando um valor de corte $c = l(x0)$ varia de uma forma contínua do seu maior valor possível até o menor possível. Este conjunto de $matrizes2 \times 2$ é único para as duas distribuições de X.

Definição de ROC em termos de x - Uma ROC sumariza o conjunto possível de $matrizes2 \times 2$, que resulta quando intervalos disjuntos do eixo do x são sucessivamente adicionados ao intervalo de aceitação, a inclusão de intervalos começa com o intervalo vazio e termina com todo o eixo do x. Os conjuntos possíveis de $matrizes2 \times 2$ estão restristos pelas duas distribuições de X.

Por exemplo, dado um par de distribuições de X contínuas, apenas uma ROC resulta da utilização de $l(x)$ como critério de decisão. Dado o mesmo par de distribuições, existe um grande número de ROC s, cada uma dependendo da ordem de inclusão dos intervalos em x no critério de aceitação [24].

O sistema de coordenadas da ROC apresenta como ordenadas a proporção de acertos, $P(S|sn)$, e como abcissas a proporção de falsos alarmes $P(S|n)$. Quando as probabilidades

são projetadas linearmente, os valores de coordenadas variam de zero até um, e todas as ROC possíveis estão limitadas por um quadrado unitário. A diagonal positiva deste quadrado é denominada linha do acaso, em que $P(S|sn) = P(S|n)$; a diagonal negativa, corresponde a $P(S|sn) = 1 - P(S|n)$.

A figura 5.4 ilustra o sistema de coordenadas utilizado para representar uma ROC. Cada ponto neste espaço ROC corresponde a uma $matriz2 \times 2$.

Se o observador utilizar uma regra de decisão pura, isto é, se o observador for coerente nas suas respostas para cada x, então, de acordo com as definicões dadas de ROC, esta deverá começar em $(0, 0)$ e terminar em $(1, 1)$; sob estas condicões a ROC deverá ser não decrescente em todo o seu percurso [24].

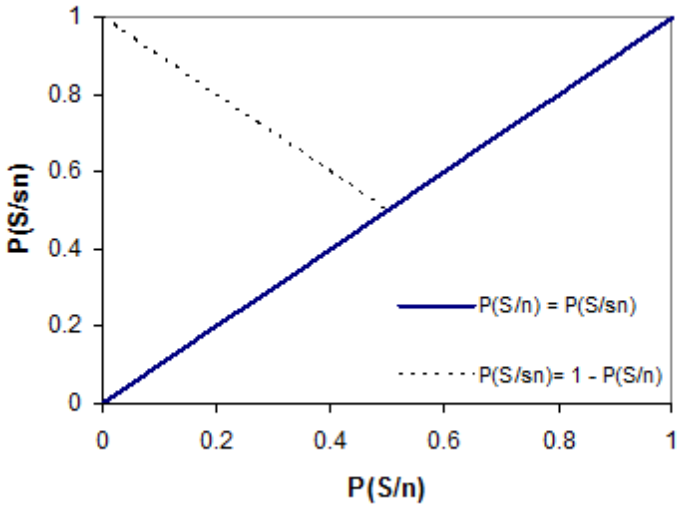


Figura 5.4: Sistema de coordenadas para representação de uma ROC

A ROC para um observador, em termos da razão de verossimilhanças, sumariza uma relaçãao específica entre as duas distribuiçõoes de probabilidade.

Considere-se a variável em estudo representada por x e que valores baixos de x favorecem a decisão "normal"$(T-)$ e valores elevados de x favorecem a decisão "fraude "$(T+)$. Designe-se ainda, por $f(x|A)$ a distribuição dos valores de x para os casos designados anormais, xA, e por $f(x|N)$ a distribuiçãodos valores de x para os casos designados

normais, xN; ou seja, a distribuição de xA deverá ser centrada à direita da de xN .

Graficamente, a situação descrita, poderia ser ilustrada pela figura 5.5.

Como se pode verificar a partir desta figura, as distribuições de xA e xN, sobrepõe-se, e isto significa que, alguns dos casos inicialmente identificados como normais poderão ter leituras como fraude, e por outro lado, alguns dos casos inicialmente identificados como fraude poderão ter leituras como normais.

Para qualquer teste de diagnóstico é fixado um valor de corte para a variável em estudo, valor que determina a classificaçãao dos indivíduos como fraudadores ou normais. Assim, qualquer teste é avaliado pela comparação relativa da fração de verdadeiros positivos (FVP), fração de falsos positivos (FFP), fração de verdadeiros negativos (FVN) e fração de falsos negativos (FFN).

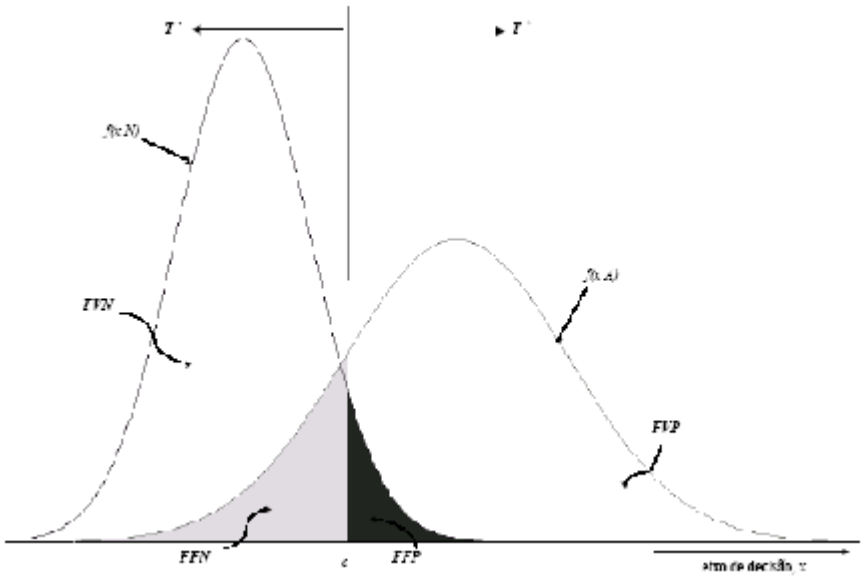


Figura 5.5: Sobreposição das distribuições hipotéticas (normais e fraudadores)

Em termos de diagnóstico, a fração de verdadeiros positivos (FVP) corresponde à probabilidade de decidir que a característica em questão está presente, quando de fato está presente. Por outro lado, a fração de verdadeiros negativos (FVN) corresponde à

probabilidade de decidir que a característica está ausente, quando de fato está ausente.

Estas duas definições conduzem a outras duas diretamente relacionadas, a fração de falsos positivos e a fração de falsos negativos, dadas por:

$$FFP = \frac{\text{número de decisões falsas positivas}}{\text{número de casos realmente negativos}}$$

e

$$FFN = \frac{\text{número de decisões falsas negativas}}{\text{número de casos realmente positivos}}$$

Note-se que estas frações representam, respectivamente, as frações de casos designados por realmente negativos e as frações de casos designados por realmente positivos que são decididos incorretamente.

Se assumirnos que todos os casos podem ser diagnosticados como positivos ou negativos (considerando determinada característica), então, o número de decisões corretas mais o número de decisões incorretas deverá ser igual ao número de casos com esse estado atual.

Assim, verifica-se que:

$$FVP + FFN = 1$$

e

$$FVN + FFP = 1$$

A figura 5.5 pretende explicitar a relação entre o valor de corte e a definição dessas razões, sendo claro que diminuir a FFP conduz a um aumento de FFN.

Em geral, um teste de diagnóstico tende a ser avaliado por duas destas medidas, FVP (sensibilidade) e FVN (especificidade). Metz [19] define sensibilidade como sendo a probabilidade de decidir se a característica em questão está presente quando de fato está presente, e especificidade como sendo a probabilidade de decidir se a característica em questão está ausente quando, de fato está ausente. Em termos de diagnóstico, pode-se definir sensibilidade como a capacidade que um teste tem para detectar a característica, e a especificidade como a capacidade que o teste tem para excluir os indivíduos isentos de característica de análise. Assim, valores de corte elevados, conduzem a um teste pouco sensível e muito específico, por outro lado, valores de corte baixos, conduzem a um teste

muito sensível e pouco específico.

Num teste de diagnóstico as hipóteses podem ser definidas como:

H_0: O indivíduo é fraudador, XA

H_1: O indivíduo é normal, XN ,

consequentemente:

$\alpha = Prob(\text{erro tipo I}) = P(\text{rej } H_0|H_0) = P(T-|XA) = 1 - P(T+|XA) = 1-$ *sensibilidade*

$\beta = Prob(\text{erro tipo II}) = P(\text{aceitar } H_0|H_1) = P(T+|XN) = 1-P(T-|XN) = 1-$ *especificidade*

Atendendo a que o valor de corte define a região de rejeição, isto é, define a dimensão dos erros de tipo I e de tipo II, à medida que se varia o valor de corte estes erros vão variando, existindo um balanço, à medida que α aumenta, β diminui, e vice-versa.

O objetivo é maximizar a sensibilidade e especificidade do teste determinando um valor de corte, fixando um par sensibilidade/especificidade. Estes pares podem ser representados como valores de coordenadas "y" e "x" dando origem ao gráfico designado por curva ROC, permitindo uma noção da capacidade de discriminação do modelo proposto.

Por definição, uma curva ROC é a representação gráfica dos pares sensibilidade ou FVP (ordenadas) e 1- especificidade ou FFP (abcissas), resultantes da variação do valor de corte ao longo de um eixo de decisão, x, e a representação gráfica assim resultante é designada por curva ROC no plano unitário.

Desta forma, uma curva ROC é uma descrição empírica da capacidade do modelo preditivo poder discriminar entre dois estados num universo, onde cada ponto da curva representa um compromisso diferente entre a FVP e a FFP que pode ser adquirido pela adoção de um diferente valor de corte de anormalidade ou nível crítico de confiançaa no processo de decisão [19].

Sob o ponto de vista da teoria de testes de hipóteses estatísticas, uma curva ROC é conceitualmente equivalente a uma curva que mostra a relação entre a potência de teste e a probabilidade de cometer um erro de tipo I com a variação do "valor crítico" (valor de corte) do teste estatístico.

Consoante os critérios adotados poder-se fazer corresponder um ponto na curva ROC. Assim, pode-se definir, um critério "estrito" (por exemplo, apenas se designa o fraudador

quando a evidência da caracterítica é muito forte) como sendo aquele que conduz a uma pequena fração de falsos positivos e também a uma relativamente pequena fração de verdadeiros positivos, isto é, gera um ponto na curva ROC que se situa no canto inferior esquerdo do espaço ROC. Progressivamente critérios menos estritos conduzem a maiores frações de ambos os tipos, isto é, pontos colocados no canto superior direito da curva no espaço ROC. Esta situação pode ser descrita graficamente pela curva ROC apresentada na figura 5.6.

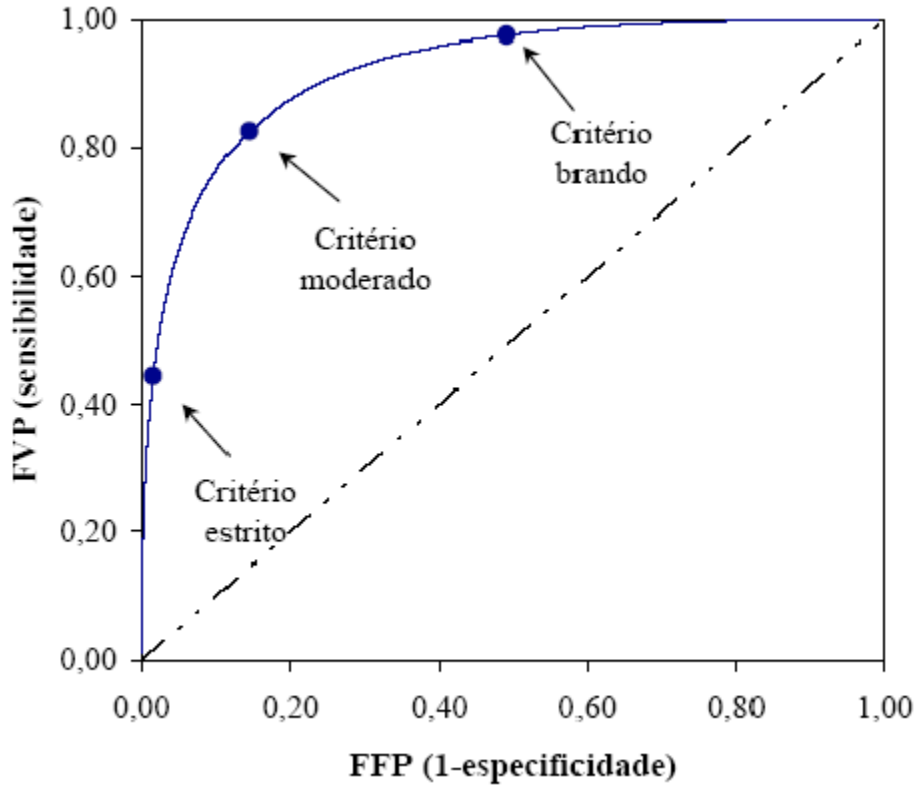


Figura 5.6: Curva ROC com variação do critério de decisão.

No que diz respeito ao desempenho de diferentes modelos preditivos, e considerando a situação em que as curvas ROC associadas a modelos preditivos não se cruzam, o modelo com a curva ROC mais próxima do canto superior esquerdo, fornece um maior poder discriminante.

Na figura 5.7 exemplificam-se três graus de discriminação possíveis fornecidos pelas curvas ROC. Quando as curvas ROC se cruzam então podem-se classificar osmodelos para um conjunto de frações de falsos positivos ou verdadeiros positivos de interesse no sentido da identificação da característica.

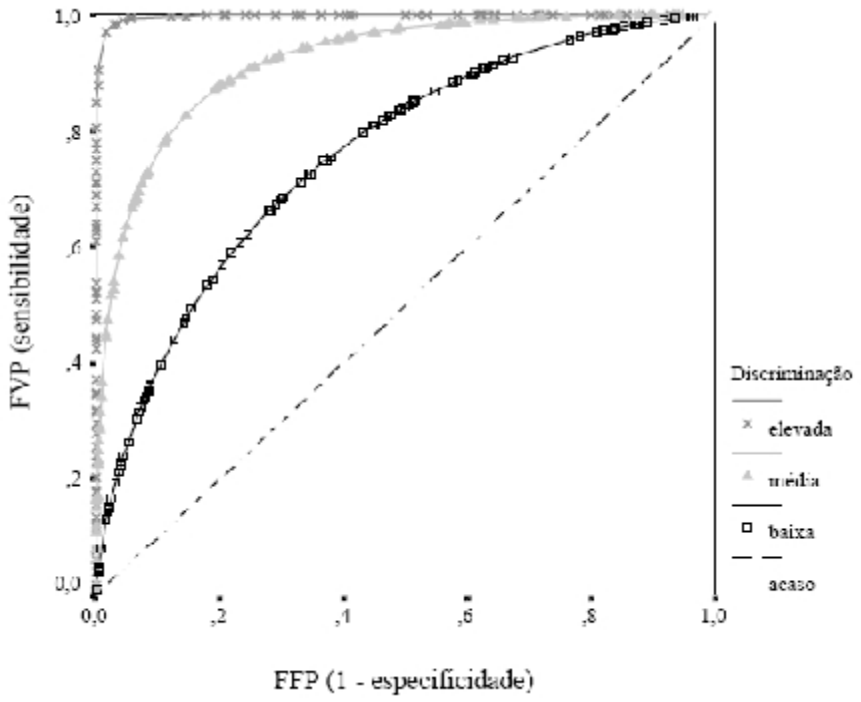


Figura 5.7: Curvas ROC para 3 graus distintos de discriminação

Capítulo 6

Procedimento Bayesiano para Análise de Fraude

O avanço das capacidades computacionais, na última década, possibilitou a implementação de métodos de aproximação numéricos (técnicas de simulação) promovendo desenvolvimento do tratamento e análise estatística de problemas cada vez mais complexos. A facilidade de calcular integrais de funções complexas e de muitas dimensões deu um novo impulso a inferência bayesiana. No caso da inferência clássica se está interessado nas propriedades dos estimadores e na distribuição amostral de estatísticas de teste.

De maneira geral pode-se dizer que um problema de inferência estatística é um problema de otimização de função ou a solução de integrais. Assim os algoritmos para solucionar esses tipo de problema não são determinísticos, isto é, baseiam-se em simulação de números "pseudo"aleatórios de uma distribuição de probabilidade. As limitações deste tipo de abordagem são o tempo computacional e a capacidade de armazenamento dos valores simulados.

Este capítulo tem por objetivo apresentar técnicas baseadas em simulação com o propósito de calcular a incerteza associada a estimadores. A incerteza ou a variabilidade dos estimadores depende muito do tamanho da amostra que é utilizada no calculo das estimativas, quando a amostra é muito pequena temos grande variabilidade associada.

Em inferência bayesiana cada problema é único e tem um contexto real próprio onde θ é uma quantidade significativa que encerra graus de conhecimento variáveis de acordo com a natureza do problema e o grau de conhecimento subjetivo do investigador sobre o

problema. Desta forma, os pesquisadores, ditos bayesianos, denominam a probabilidade que capta esta variabilidade da informação subjetiva de distribuição a *priori*.

6.1 Análise considerando Priori não Informativa

6.1.1 Técnica de Monte Carlo via cadeias de Markov

Nos métodos de simulação não iterativos gera-se uma amostra da distribuição de interesse em um único passo. Os valores são gerados de maneira idependente e não há preocupação com a convergência do algoritmo, tendo como condição nescessária um tamanho de amostra suficientemente grande. Para problemas complexos a alternativa é a utilização de técnicas iterativas - Monte Carlo via cadeias de Markov (MCMC), onde não existe a preocupação em encontrar uma densidade de importância que seja, simultaneamente, uma boa aproximação da densidade de interesse e fácil de ser amostrada.

Neste tipo de abordagem obtém-se amostras da distribuição de interesse e calcula-se estimativas amostrais de características desta distribuição. A principal diferença entre os métodos não iterativos e iterativos, baseados em cadeias de Markov, é que nestes os valores gerados não são independentes [25],[26].

O objetivo é simular um passeio aleatório no espaço x, que converge para uma distribuição estacionária, que é a distribuição de interesse do problema. Uma cadeia de Markov é um processo estacionário $\{X_0, X_1, \ldots\}$, onde a distruição de X_t dado os valores anteriores $X_0, X_1, \ldots, X_{t-1}$ depende apenas de X_{t-1}:

$$P(X_t \in A | X_0, X_1, \ldots, X_{t-1}) = P(X_t \in A | X_{t-1}) \tag{6.1}$$

para qualquer subconjunto A. Os métodos MCMC requerem que a cadeia seja:

- homogênea: isto é que as probabilidades de transição de um estado para outro sejam invariantes;
- irredutíveis: onde cada passo pode ser atingido de qualquer outro em um número infinito de iterações;
- aperiódica: isto é, que não haja estados absorventes.

Seja uma distribuição $\pi(x), x \in \mathbb{R}^d$, conhecida a menos de uma constante multiplicativa mas complexa o suficiente para não ser possível obter uma amostra diretamente. Dadas as realizações $\{X_t, t = 0, 1, \ldots\}$ de uma cadeia de Markov e π a distribuição de equilíbrio:

$$X^t \overset{t\to\infty}{\longrightarrow} \pi(x) \quad e \quad \frac{1}{n}\sum_{t=1}^{n} g(X_i^t) \overset{n\to\infty}{\longrightarrow} E_\pi\left[g\left(x\right)\right] \tag{6.2}$$

Assim a média aritmética é um estimador consitente da média teórica. vale ressaltar que os valores iniciais influenciam o comportamento da cadeia, mas a medida que as iterações aumentam a cedia converge para uma distribuição de equilíbrio. Na prática os valores iniciais são descartados (burn in).

Os algoritmos de Metropolis-Hastings utilizam a abordagem dos métodos de rejeição, isto é, um valor é gerado a partir de uma distribuição auxiliar e aceito com uma dada probabilidade, garantindo a convergência da cadeia para a distribuição de equilíbrio.

Seja uma cadeia no estado x e um valor x' é gerado de uma distribuição $q(.|x)$. O novo valor de x'é aceito com probabilidade:

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right\} \tag{6.3}$$

onde π éa distribuição de interesse. Na abordagem bayesiana a distribuição de interesse é a própria posterior $\pi(\theta|x)$ e a probabilidade de aceitação assume a forma:

$$\alpha(\theta, \theta^*) = \min\left\{1, \frac{p(x|\theta^*)}{p(x|\theta)}\frac{p(\theta^*)q(\theta|\theta^*)}{p(\theta)q(\theta^*|\theta)}\right\} \tag{6.4}$$

6.1.2 Estimação dos Hiperpâmetros da Regressão Logística via Metropolis - Hastings

Supondo dois grupos G_1e G_2 (fraude e não fraude) e que se quer estimar um modelo estatístico para classificar um novo indivíduo num dos grupos com base num vetor de p covariadas $x = (x_1, \ldots, x_{p-1})$. O modelo logísitico considera que a probabilidade de um indivíduo pertencer ao grupo $G_j, \quad j = 1, 2$ dado o vetor de covariadas assume a seguinte forma:

$$P(G_1|x,\beta) = \frac{\exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)} \tag{6.5}$$

$$P(G_2|x,\beta) = 1 - P(G_1|x,\beta) = \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)} \tag{6.6}$$

onde $\beta = \left(\beta_0, \ldots, \beta_{p-1}\right)$ é um vetor de p parâmetros reais. A regra desta abordagem é classificar o indivíduo em G_1 se $P(G_1|x) > 1/2$ o que equivale a dizer que $(\beta_0 + \sum_{i=1}^{p} \beta_i x_i) > 0$. Classifica-se o indivíduo com G_2 caso esta condição não seja satisfeita. Na abordagem clássica os parâmetros são estimados pelo método de máxima verossimilhança utilizando os dados de indivíduos corretamente classificados, e q regra é utilizada substuindo-se os parâmatros pelas suas estimativas. O que é equivalente a supor que, para cada indivíduo, a variável indicadora do grupo a que o indivíduo pertence y é uma variável de Bernoulli com parâmetro $p(x) = P(G_1|x,\beta)$. A verossimilhança relativa a amostra é:

$$L(\beta|D) \propto \prod_{i=1}^{n} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1-y_i} \tag{6.7}$$

onde D representa a informação relativa à experiencia com n indivúdos, $D = \{y_i, x_i, i = 1, \ldots, n\}$. Abordando o problema do ponto de vista bayesiano considera-se a distribuição a priori $h(\beta)$ para os parâmetros do modelo e obtém-se a distribuição a posteriori

$$h(\beta|D) \propto \exp\left\{\sum_{i=1}^{n}(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)\right\} \prod_{i=1}^{n} \left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right)^{-1} h(\beta) \tag{6.8}$$

O algoritmo metropolis-hastings é utilizado quando a posteriori a ser amostrada, ou seja, as densidades condicionais apresentam distribuições que não são conhecidas.

Seja $q(\theta, \theta^*)$ uma transição, isto é, uma probabilidade do movimento de θ para θ^*. Geralmente o processo se move de θ para θ^* com mais freqüência do que de θ^* para θ violando a condição de reversibilidade

$$\pi(\theta) q(\theta, \theta^*) = \pi(\theta^*) q(\theta^*, \theta). \tag{6.9}$$

Capítulo 7

Modelo de Propensão à Fraude: etapas do desenvolvimento

O processo de análise de fraude utiliza diversas técnicas de análise de dados com tecnologias de alto padrão, suportando as tomadas de decisões das diversas áreas gerenciais de uma empresa. As etapas deste processo são:

- Definição do Problema ;
- Identificação e amostragem dos dados;
- Definição da Modelagem;
- Treinamento, validação e teste dos modelos;
- Escolha do modelo que melhor descreve o problema ;
- Interpretação dos resultados analíticos oriundos da modelagem;
- Implementação dos resultados da modelagem na base de clientes.

7.1 Objetivo Principal

O Gerenciamento de Fraudes tem por objetivo detectar, prevenir e minimizar eventos de fraude em Operadoras, através do desenvolvimento de modelos preditivos para a identificação de padrões suspeitos.

O objetivo é que todo conhecimento sobre as ações fraudulentas seja gerado e transferido a empresa operadora em tempo real ou quase real.

7.2 Objetivos Específicos

O gerenciamento de fraudes em telecomunicações tem como objetivos específicos:

- Detectar uma chamada telefônica fraudulenta;
- Quantificar a chance desta chamada ser fraudulenta;
- Caracterizar os perfis dos fraudadores;
- Identificar nichos de clientes com maior propensão à fraude;
- Gerar conhecimento sobre as ações fraudulentas, orientando a empresa operadora no combate à fraude; e
- Caracterizar novos tipos de fraudes.

7.3 Considerações

No gerenciamento da ocorrência de fraude, deve-se levar em conta os dois tipos de erro envolvidos nesta análise. Aquele que mais preocupa os analistas é o risco de não se classificar um evento fraudulento como tal. No entanto, é importante que se considere também o prejuízo para a operadora quando se comete o erro de classificar uma ligação honesta como fraudulenta, pois no cenário competitivo atual, um cliente injustamente abordado/penalizado pode facilmente migrar para uma empresa operadora concorrente, ocasionando mais perda de receita para a operadora.

Atualmente muitos fraudadores se beneficiam das fragilidades de muitas redes telefônicas para aplicar fraudes de voz , já que em alguns casos somente após dois ciclos de faturamento as operadoras percebem as ações dos fraudadores sobre suas redes.

7.4 Estratégias de Modelagem

As estratégias de modelagem podem ser classificadas em duas categorias: aprendizado supervisionado e aprendizado não supervisionado. Os métodos de aprendizado supervisionados são desenvolvidos quando existe uma variável resposta (ou alvo) com valores observados, cujas predições usam os valores das variáveis explicativas presentes na base de dados. Os métodos de aprendizado não supervisionado tendem a serem desenvolvidos para uma variável resposta que não possui valores observados e cujas predições também se baseiam nos valores das variáveis explicativas.

A Tabela 1 mapeia as técnicas de modelagem[1] pelo objetivo da modelagem.

Objetivo	Supervisonados	Não Supervisonados
Predição	Modelos de Regressão Redes Neurais Árvores de Decisão	não se aplica
Classificação	Análise de Discriminante Redes Neurais Árvores de Decição	Clustering Redes Neurais Mapas Auto Organizáveis
Exploração	Árvores de Decição	Componentes Principais Clustering
Associação	não se aplica	Análise Fatorial Sequência de Associações

É possível que uma empresa esteja realizando pela primeira vez uma análise de fraude, neste caso, a variável resposta dicotômica (ser ou não fraude) ou contínua (propensão à fraude) não tem registros conhecidos a priori, o que exige a utilização de métodos não supervisionados como os descritos na Tabela 1. Estes métodos organizam a base de dados em grupos (segmentos) conforme as similaridades dos dados e servem para descobrir padrões até então desconhecidos. Os resultados destes procedimentos devem ser testados antes de aplicados sobre toda a base de dados.

Os métodos não supervisionados disponibilizam mais informações sobre os perfis de

[1] *Using Data Mining Techniques for Fraud Detection. SAS Intitute, 1999*

fraudes e dos fraudadores que os tradicionais relatórios técnicos utilizados pelas empresas para descrição de fraudes em seu sistema. Além disso, provêem informações que servirão de input para a construção de uma base de conhecimento para predizer mais rapidamente uma fraude.

Nos métodos supervisionados de Mining, a amostra aleatória deve ser formada por registros de fraudadores e não fraudadores a fim de que os modelos construídos permitam a classificação de um novo registro em um dos grupos (fraudador ou não fraudador). Desta forma, avalia-se quão preciso é o modelo para classificar um novo registro.

A Fraude por subscrição é caracterizada pela obtenção da assinatura de um serviço com a intenção de fraude. Este tipo de fraude não é específico de tecnologia, mas ocorre nos processos de ativação normais de uma operadora e é analisado através de informações no cadastramento.

O objetivo é identificar os padrões de fraude utilizados em ativação cadastral com o objetivo de fraudar a operadora de Telecomunicações. A análise da fraude por subscrição deverá ter duas abordagens:

- Identificação da característica de fraude na inclusão do prospect na base de clientes da empresa.
- Identificação de padrão suspeito de fraude em clientes novos, com pouco tempo de relacionamento com a operadora (até 3 meses).

A concorrência acirrada entre as operadoras de telefonia exige esforços cada vez maiores por parte das empresas por uma fatia maior de participação no mercado, aumentando a carteira de clientes. Assim, o objetivo desta modelagem é identificar os clientes que apresentam alta propensão à fraude por subscrição, subsidiando as áreas de negócio a evitarem que isso ocorra efetivamente. A capacidade de anteceder as ações dos possíveis fraudadores visa reduzir a fuga de receita, uma vez que as ações preventivas estarão sendo alocadas cada vez mais acertadamente.

A figura 7.1 detalha a análise e gerenciamento da fraude por subscrição como um processo contínuo, iniciando-se no momento de ativação do serviço até completar-se 3 meses de relacionamento. Neste intervalo de tempo o comportamento do novo cliente é analisado minuciosamente. Esta análise será possível através de um encadeamento de

modelos estatísticos, este procedimento visa maximizar a sensibilidade e especificidade do processo de detecção da fraude.

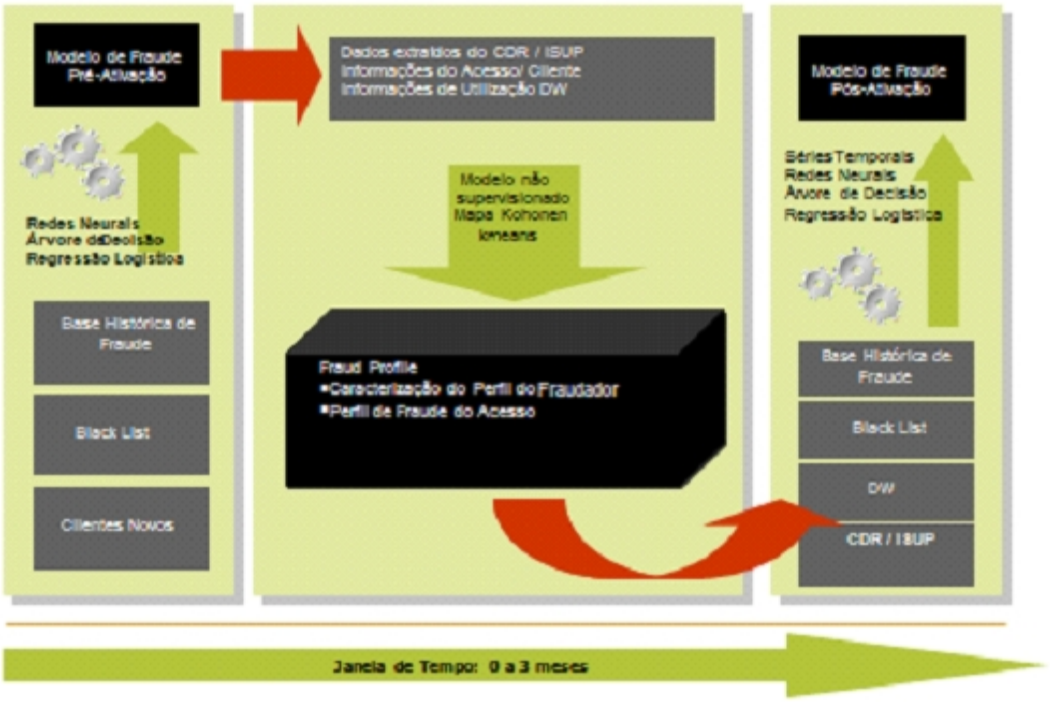


Figura 7.1: Arquitetura Modelagem Fraude de Subscrição

A aplicação abordada neste estudo compreende o modelo de fraude pré-ativação, cujo objetivo é predizer (classificar) a propensão à fraude antes da habilitação do aparelho celular, ou tão logo seja factível.

Os 3 primeiros dias de utilização da nova linha telefônica são fundamentais para caracterizar o fraudador. Neste intervalo de tempo avalia-se o perfil de comportamento de utilização do serviço dos novos clientes em comparação com os clientes idôneos. Alterações cadastrais do assinante logo após a ativação do sistema podem ser um indicativo de má fé do novo cliente.

O modelo com maior aderência e performance na classificação do evento de fraude foi o modelo baseado em regressão logística. Nesta etapa utilizamos uma amostra de 7950 registros, com 2450 registros identificados como fraude.

O planejamento macro do processo de construção da base amostral que será utilizada para a elaboração do modelo de fraude pré-ativação é apresentado na figura abaixo.

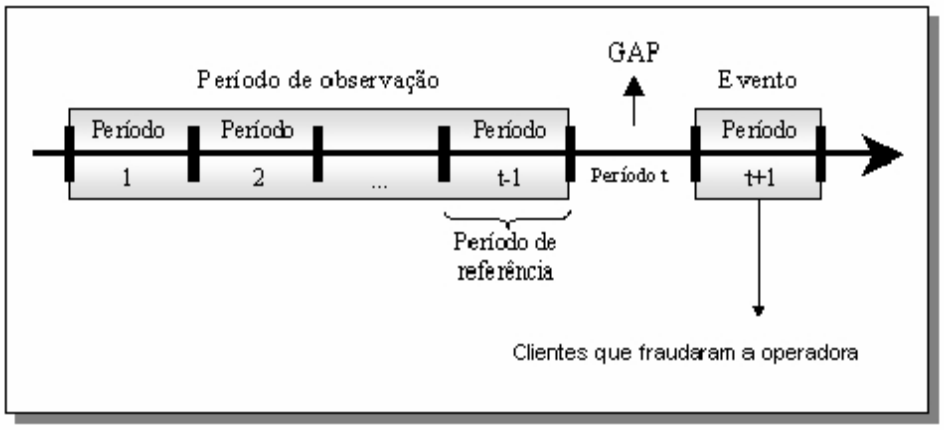


Figura 7.2: Janela de Tempo de Análise do modelo de Fraude de Subscrição

A figura 7.2 descreve os períodos de observação de dados históricos de fraude, de latência de tempo (GAP) e de estimação do evento de interesse. O período de observação refere-se na figura à base histórica de fraude (Período 1, Período 2,..... , Período$_{t-1}$). O período de latência (GAP) é a fase de desenvolvimento do modelo de fraude de pré-ativação e o período de estimação do evento (Período$_{t+1}$) corresponde ao instante para o qual o modelo de fraude de pré-ativação estima a probabilidade do novo cliente vir a cometer fraude contra a empresa operadora.

O planejamento proposto acima pode sofrer ainda algumas alterações (ou adaptações) em decorrência, por exemplo, da baixa quantidade de eventos encontrados ou até mesmo em função da operacionalização envolvida no processo.

Por apresentar uma regra clara para definir a variável resposta, o modelo de preditivo de fraude pré-ativação foi tratado através de técnicas supervisionadas de modelagem de dados (Decision Tree, Regressão Logística e Redes Neurais).

7.4.1 Conjunto de Dados e Pré-processamento

As informações para a construção dos modelos de Mining para o Gerenciamento de Fraude de Subscrição foram simuladas considerando características de fraude observadas no mercado de Telecomunicações Brasileiro, aliada às características de utilização do serviço móvel celular observado em estudos anteriormente desenvolvidos. A simulação do conjunto de dados de Pré-ativação foi realizada conforme o Guia de Variáveis:

- id_acesso: Identificação do Acesso
- plano_servico: Plano de Serviço (1..n, quantos forem comercializados pela empresa)
- ativ_promocao: Ativação em Promoção (0 não se aplica, 1...n, quantas forem as promoções realizadas pela empresa)
- vig_contrato: Vigência do Contrato (0 não se aplica, 6 meses, 12 meses, 18 meses, 24 meses)
- tipo_contrato: Tipo de Contrato (0 Pré Pago, 1 Pós Pago Renovação automática, 2 Pós pago Tempo Fixo, 3 Pós Pago Indeterminado)
- mes_inicio_contrato: Mês de início de contrato
- hab_chamada_inter: Habilitação para chamada Internacional
- modelo_handset: Modelo do handset
- marca_handset: Marca do handset
- handset_usado: Handset usado (0: não; 1:sim)
- serial_handset_usado: Serial do Handset usado
- ocupacao: Ocupação
- sexo: Sexo
- idade: Idade
- faixa_etaria: Faixa Etária

- canal_venda: Canal de venda
- dealer_autorizada: Dealer / Autorizada
- cep: CEP
- dig_regiao_fiscal: Dígito verificador da região fiscal do CPF
- ind_modelo: Indicador de Fraude por Modelo de Handset
- ind_marca : Indicador de Fraude por Marca de Handset
- ind_promocao: Indicador de Fraude por Promoção
- ind_faixa_etaria: Indicador de Fraude por Faixa Etária
- ind_plano_servico: Indicador de Fraude por Plano de Serviço
- ind_cep: Indicador de Fraude por CEP
- ind_sexo: Indicador de Fraude por Sexo
- ind_mes_assinatura: Indicador de Fraude por Mês de Assinatura
- ind_canal_vendas: Indicador de Fraude por Canal de Vendas
- ind_dealer_autorizada: Indicador de Fraude por Dealer / Autorizada
- ind_faixa_renda: Indicador de Fraude por Faixa de Renda
- ind_regiao_fiscal: Indicador de Fraude por último dígito do CPF (referente a região fiscal de registro de CPF)
- flg_cpf_assoc_fraude: Flag de CPF associado a fraude
- flg_cep_assoc_fraude: Flag de CEP associado a fraude
- flg_esn_assoc_fraude: Flag de ESN associado a fraude
- restricao_cpf_serasa: Restrição do CPF no SERASA
- status_cpf_rec_federal: Status de CPF na Receita Federal

- credit_score: Credit Score
- faixa_renda: Faixa de Renda
- flg_fraude: Flag de Fraude

A técnica de *oversampling* é um dos métodos propostos pelos pesquisadores da área para solucionar o problema de eventos raros, balanceando artificialmente a distribuição das classes (evento e não-evento) no conjunto de dados. Muitos sistemas de aprendizado assumem que as classes estão balanceadas. Quando isso não acontece, ou seja, as classes são desbalanceadas, esses sistemas acabam induzindo um classificador incapaz de predizer a classe minoritária com precisão razoável. Por esse motivo, para utilizar tais sistemas, usamos o artifício de *oversampling*, que visa balancear a distribuição das classes por meio da super estimação da incidência da classe minoritária.

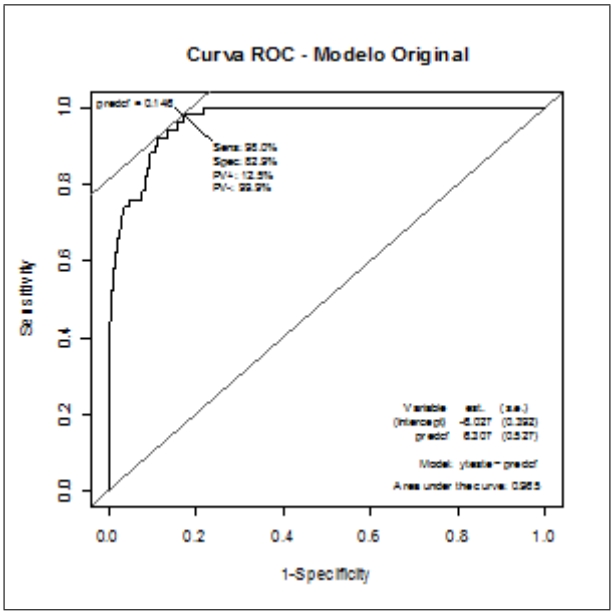
Com o objetivo de se obter amostras de treinamento e teste, a base de dados foi sub dividida em : treinamento e teste. A técnica de *oversampling* foi utilizada apenas na base de treinamento.

A base de teste é utilizada para avaliar a validade ou acurácia do modelo. De modo geral, este conjunto amostral tem por objetivo avaliar a performance do modelo estimado.

O modelo de fraude pré-ativação foi elaborado considerando uma base treinamento de 7950 clientes, com 2450 casos de fraude. O modelo de regressão logística final foi estimado pelo procedimento Stepwise, identificando-se o conjunto de variáveis significativas para classificar os clientes em fraudadores e não fraudadores, de acordo com a tabela abaixo:

	Variável	Parâmetro	σ	z	$P(>\|z\|)$
	Intercepto	-9.527132	0.583260	-16.334	$< 2e^{-16}$
x_1	hab_cham	0.995613	0.188465	5.283	$1.27e^{-07}$
x_2	modelo_h	-0.023685	0.011253	-2.105	0.035315
x_3	handset	0.909615	0.169599	5.363	$8.17e^{-08}$
x_4	dealer	-0.016478	0.004249	-3.878	0.000105
x_5	cep	-0.048147	0.002379	-20.236	$< 2e^{-16}$
x_6	ind_marc	2.234367	0.973855	2.294	0.02177
x_7	ind_prom	6.162027	1.024243	6.016	$1.79e^{-09}$
x_8	ind_faix	4.040227	0.255636	15.805	$< 2e^{-16}$
x_9	ind_plan	4.119263	0.396115	10399	$< 2e^{-16}$
x_{10}	ind_sexo	1.877031	0.231179	8.119	$< 2e^{-16}$
x_{11}	ind_cana	3.082391	0.258695	11.915	$6.98e^{-10}$
x_{12}	flg_cpf	2.827465	0.278995	10.134	$< 2e^{-16}$
x_{13}	flg_cep	3.770656	0.208562	18.079	$< 2e^{-16}$
x_{14}	flg_esn	1.838004	0.298065	6.166	$6.98e^{-10}$
x_{15}	restrica	0.981221	0.107939	9.091	$< 2e^{-16}$
x_{16}	status_c	0.682464	0.061452	11.106	$< 2e^{-16}$
x_{17}	credit_s	11.959379	0.408440	29.281	$< 2e^{-16}$

A curva ROC é uma métrica de avaliação de qualidade do modelo, que consiste, basicamente em observar a taxa de verdadeiro positivo (sensitividade) versus a taxa de falso positivo (1 - especificidade). A eficiência do modelo depende de quão bem este consegue separar o grupo que experimentou o evento daqueles que não experimentaram o evento. A acurácia é medida pelo valor da área entre a curva ROC e a diagonal principal (a diagonal principal representa o pior dos casos, ou acurácia zero). O intervalo de variação dessa área é de 0,5 (acurácia zero) a 1,0 (separação perfeita dos casos).



Como a idéia principal é maximizar a sensitividade e ao mesmo tempo minimizar o número de casos falso positivos, o gráfico auxilia também na definição do ponto de corte ótimo, que visa obter a maior quantidade de predições corretas (diagonal da matriz de confusão), no caso do modelo analisado o ponto de corte recomendado é 0.146, que produz a seguinte tabela

	Fraude		
Modelo	sim	não	Total
Fraude	49	342	391
não Fraude	1	1658	1659
Total	50	2000	2050

A sensibilidade do modelo é de 98%, isto é, consegue classificar corretamente 98% dos casos de fraude. A especificidade é de 82.9%, em outras palavras 82.9% dos não fraudadores são classificados corretamente. No entanto a quantidade de casos de clientes "bons "classificados como fraudadores demanda um trabalho minucioso das áreas de Análise de Fraude, em detrimento ao trabalho investigativo de novos padrões de fraude, além de possibiltarem o surgimento de atrito entre a operadora e os clientes considerados "bons". O objetivo é ter um modelo capaz de minimizar este tipo de situação, minimizando o erro dos falsos positivos.

Capítulo 8

Aplicação

Neste capítulo iremos demonstrar as metodologias de análise clássica e bayesiana aplicadas no conjunto de dados reais de um problema de Fraude de subscrição de uma Empresa de Telecomunicações celular.

Este estudo foi divido em dois estágios, na primeira fase foi realizada um estudo que permitiu a comparação entre a metodologia clássica para estimação de um modelos logístico e metodologia bayesiana com prioris não informativas. Para este objetivo foi elaborada uma estratégia amostral da base original de treinamento demonstrada na figura 8.1.

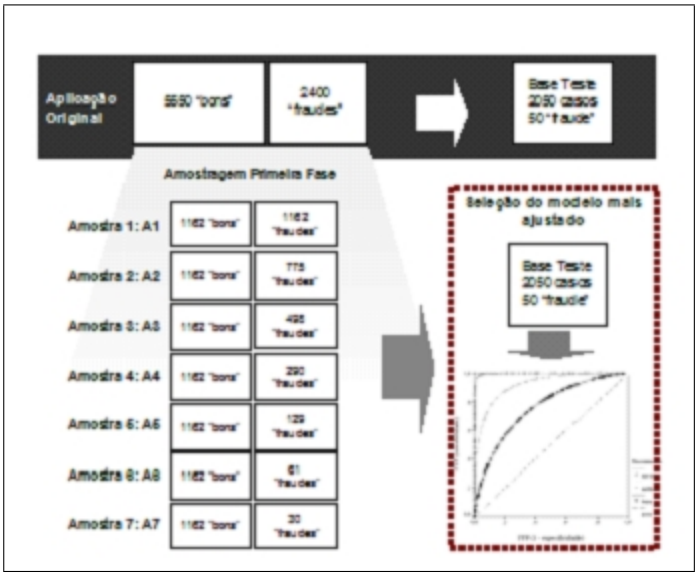


Figura 8.1: Estratégia de Amostragem 1ª fase

Esta estratégia foi adotada pois pretende-se reproduzir as dificuldades encontradas ne modelagem real de um problema de fraude de subscrição, onde a quantidade de *bons* clientes supera em muito a quantidade de *fraudadores* (a incidência de casos de fraude de subscrição é de aproximadamente 2,5%). Desta forma estruturou-se amostras com proporções balanceadas de clientes *fraudadores.*

8.1 Descrição do Problema

A base de dados deste trabalho constituiu-se da base amostral (subdividida em treinamento e teste) utilizada para modelar a propensão à fraude de subscrição em Empresa de telecomunicações Celular. Foram coletados 12 amostras de safras (um ano) de clientes *bons* que entraram na base cadastral da empresa e todos os clientes *fraudadores* que entraram no mesmo período. A base treinamento, utilizada para selecionar e estimar os parâmetros do modelo com melhor performance, foi estruturada de forma a apresentar um *oversampling* 30/70, isto é, amostra balanceada com a incidência de 30% do total de casos como clientes *fraudadores* r

O principal objetivo de comparar a performance da modelagem baseada em inferência bayesiana é extrair informações / padrões de comportamento dos eventos ditos raros, com uma incidência muito baixa na população, mas, que no caso de fraude, ocasionam perdas financeiras consideráveis para as empresas de Telecomunicações Celures.

Desta forma a ocorrência de amostras com pouca representatividade de casos de fraude é uma situação comum numa empresa de telecomunicações, e a espera para completar um número de safras superior a 3 meses pode gerar perdas financeiras significativas e misturar padrões de comportamentos distintos, uma vez que o fenômeno "fraude"está em constante mudança e reciclagem.

Para comparar o ajuste do modelo utilizando-se inferência clássica e bayesiana, foi utilizada a mesma estrutura do modelo estimado originalmente com 17 variáveis, conforme descrito no capítulo 6. Foi utilizada a técnica de *oversampling* para garantir a capacidade preditiva do modelo estimado, utilizando amostras balanceadas artificialmente para avaliar a influência na modelagem e uma amostra com a incidência observada de fraude (2,5%). A tabela a seguir descreve as amostras utilizadas.

Amostras	Casos de Fraude	Casos de não Fraude	Total de Casos	Incidência de Fraude
A_1	1162	1162	2324	50%
A_2	775	1162	1937	40%
A_3	498	1162	1660	30%
A_4	290	1162	1452	20%
A_5	129	1162	1291	10%
A_6	61	1162	1223	5%
A_7	30	1162	1192	2.5%

Para avaliar a acurácia da estimação dos parâmetros da regressão logística via Algorimo Metropolis-Hastings, utilizou-se o pacote MCMCpack do R, na primeira fase com prioris não informativas.

No caso da modelagem bayesiana para todas as amostras acima relacionadas foram utilizadas as seguintes premissas: 100.000 amostras, *burn in* de 10.000 (primeiras iterações) e lag=10. Os valores estimados dos parâmetros estão bem próximos dos valores estimados pela técnica de máxima verosssimilhança, no entanto observa-se que o Algoritmo Metropolis-Hastings fornece valores estimados com menor dispersão. Essa hipótese é comprovada ao se trabalhar com uma amostra de tamanho menor .As variáveis consideradas na modelagem bayesiana são as mesmas utilizadas na modelagem clássica.

8.2 Primeira fase: Comparação da Abordagem Clássica e Bayesiana

Esta seção apresenta os resultados obtidos a partir das amostras de treinamento citados na seção anterior, sendo utilizada a metodologia de Regressão Logística baseada em inferência clássica e bayesiana. Tanto na abordagem bayesiana quanto na abordagem clássica foram utilizadas as mesmas variáveis utilizadas no modelo original descrito no capítulo 7, onde foram testados vários modelos utilizados em Data Mining e o modelo que mais se adequou ao evento analisado foi Regressão Logística, onde utilizamos o método *Stepwise* para a

seleção de variáveis.

Nesta fase, onde o objetivo é comparar a performance da abordagem clássica e bayesiana a Regressão Logística foi utilizada em todas as amostras de treinamento, com a variável resposta identificando o cliente classificado como *fraudador* ou *não fraudador* numa série histórica dos 12 meses de captação de novos clientes da Operadora de Telecomunicações Celular.

Os modelos de Regressão logística fornecem escores de propensão à Fraude, onde quanto maior o valor do escore obtido pelo cliente, maior a chance do cliente ser classificado como *fraudador,* uma vez que o histórico de clientes fraudadores foi identificado como evento de interesse na modelagem. Desta forma os escores resultantes do modelo estão diretamente relacionados com o risco de fraude.

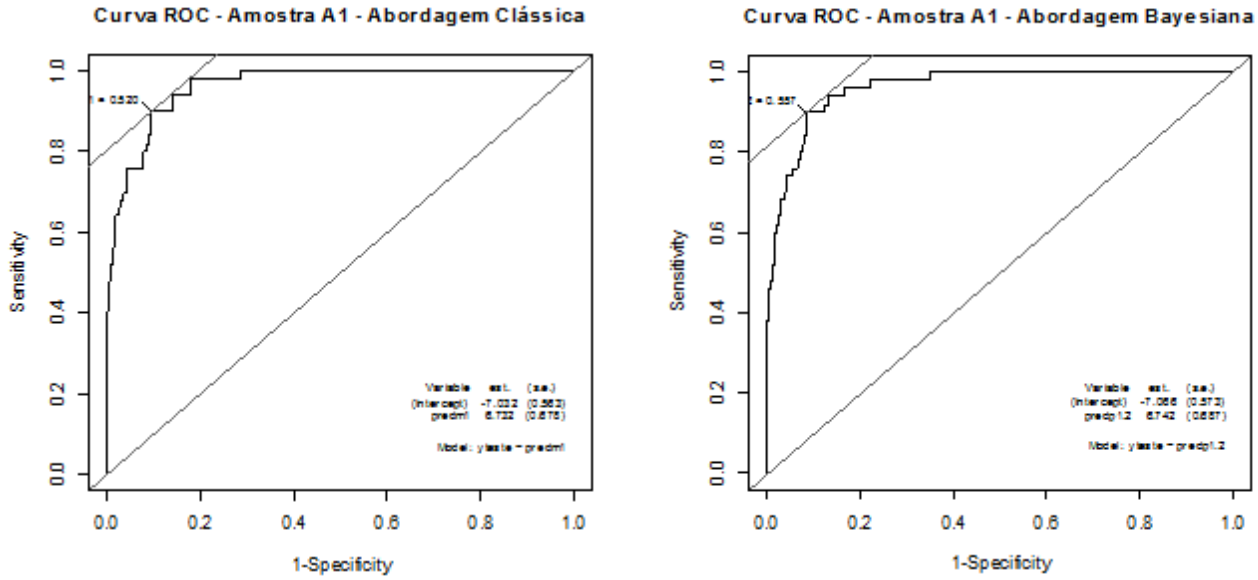
As seções subsequentes comparam os modelos estimados pela inferência clássica e inferência bayesiana, assim como a validação dos parâmetros através da Curva ROC e as medidas de capacidade de acerto (*sensibilidade* e *especificidade*) considerando como ponto de corte o valor indicado pelo índice de separação Kolgomorov-Smirnov.

8.2.1 Amostra Balanceada A1 (50/50: 50% *fraudadores* x 50% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* O resultado da modelagem clássica e bayesiana está demonstrada na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-8.963573	0.892573	-10.042	< 2e-16	-8.46032	0.877180
x_1	hab_cham	1.318896	0.326253	4.043	5.29e-05	1.33257	0.325575
x_3	handset	1.183556	0.289218	4.092	4.27e-05	1.18515	0.304254
x_5	cep	-0.049026	0.004259	-11.512	< 2e-16	-0.04963	0.004351
x_6	ind_marc	3.545881	1.139281	3.112	0.001856	-	-
x_7	ind_prom	6.520996	1.807200	3.608	0.000308	6.58746	1.854910
x_8	ind_faix	3.448993	0.416740	8.276	< 2e-16	3.54630	0.426377
x_9	ind_plan	3.900096	0.699594	5.575	2.48e-08	3.98046	0.699310
x_{10}	ind_sexo	1.847677	0.400106	4.618	3.88e-06	1.85787	0.404154
x_{11}	ind_cana	2.176902	0.312358	6.969	3.19e-12	2.21032	0.318915
x_{12}	flg_cpf	2.281248	0.508776	4.484	7.33e-06	2.35171	0.547276
x_{13}	flg_cep	3.755373	0.450405	8.338	< 2e-16	3.89720	0.461626
x_{14}	flg_esn	1.256287	0.519833	2.417	0.015661	1.32188	0.543768
x_{15}	restrica	1.158496	0.197365	5.870	4.36e-09	1.16805	0.202871
x_{16}	status_c	0.771776	0.111113	6.946	3.76e-12	0.78274	0.108741
x_{17}	credit_s	12.555106	0.733188	17.124	< 2e-16	12.88012	0.728334

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperávámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. No entanto a modelagem bayesiana produziu um modelo mais econômico, com menos variáveis associadas ao evento e com maior especificidade, isto é, classificou mais corretamente os clientes não fraudadores, minimizando o número de casos a serem acompanhados pela área de Análise de Fraude.



Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a especificidade muito próxima da sensibilidade, e um poder de discriminação muito maior com 91% de acerto. Uma outra informação a observar é que o ponto de corte na abordagem bayesiana tende a ser maior do que na abordagem clássica, em outras palavras um cliente identificado como fraudador obtem escores maiores na modelagem bayesiana.

Na tabela abaixo classificamos os clientes como fraudadores e bons de acordo com o ponto de corte estimado pela Curva ROC. Obsserva-se que a performance da aboragem bayesiana é mais eficiente nesta amostra ao identificar corretamenta clientes fraudadores e bons.

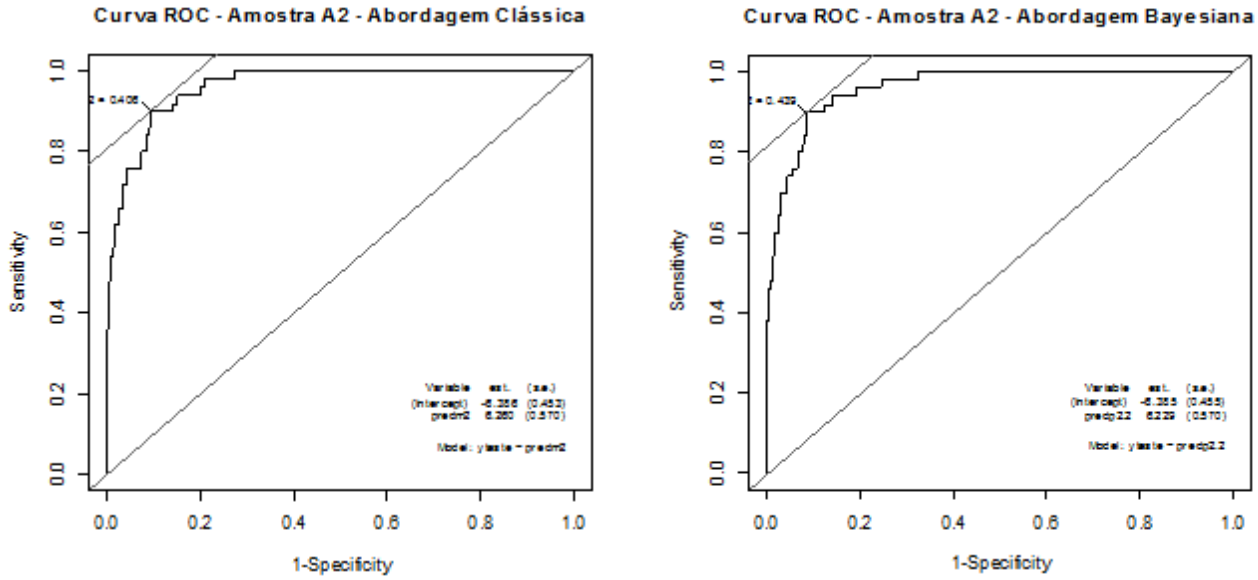
Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	sim	não	Total
	Fraude	44	187	233	Fraude	45	170	215
	não Fraude	6	1813	1819	não Fraude	5	1830	1835
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	88%				90%			
Especificidade	90.6%				91.5%			
Acurácia	90.6%				91.5%			

8.2.2 Amostra Balanceada A2 (40/60: 40% *fraudadores* x 60% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 40% de *clientes fraudadores* e 60% de *clientes bons.* O resultado da modelagem clássica e bayesiana estão demonstrados na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-10.261048	1.006440	-10.195	< 2e-16	-9.86219	0.97967
x_1	hab_cham	1.366436	0.370009	3.693	0.000222	1.39710	0.38420
x_3	handset	1.405983	0.321329	4.376	1.21e-05	1.38884	0.32497
x_5	cep	-0.047625	0.004764	-9.997	< 2e-16	-0.04849	0.00465
x_6	ind_marc	3.009603	1.281913	2.348	0.018887	-	-
x_7	ind_prom	7.560976	2.082725	3.630	0.000283	7.63756	2.07749
x_8	ind_faix	3.575340	0.478112	7.478	7.54e-14	3.63205	0.50131
x_9	ind_plan	4.515063	0.774196	5.832	5.48e-09	4.65930	0.77841
x_{10}	ind_sexo	1.828148	0.451059	4.053	5.06e-05	1.81847	0.45754
x_{11}	ind_cana	2.607733	0.356282	7.319	2.49e-13	2.64412	0.35997
x_{12}	flg_cpf	2.398325	0.559502	4.287	1.81e-05	2.46873	0.58930
x_{13}	flg_cep	3.960915	0.477507	8.295	< 2e-16	4.10424	0.49284
x_{14}	flg_esn	1.611637	0.555413	2.902	0.003712	1.68803	0.57388
x_{15}	restrica	1.281163	0.219112	5.847	5.00e-09	1.28523	0.21876
x_{16}	status_c	0.767559	0.124323	6.174	6.66e-10	0.78346	0.12719
x_{17}	credit_s	12.769013	0.826908	15.442	< 2e-16	13.14898	0.86183

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperávámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. No entanto a modelagem bayesiana produziu um modelo mais econômico, com menos variáveis associadas ao evento e com maior especificidade, isto é, classificou mais corretamente os clientes não fraudadores, minimizando o número de casos a serem acompanhados pela área de Análise de Fraude.



Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a especificidade mais alta, com uma melhor classificação de clientes não fraudadores, mimizando o grau de confundimento com os clientes fraudadores. O grau de assertividade do modelo bayesiano se manteve apesar da diminuição de casos de fraude utilizados para estimar o modelo.

Na tabela abaixo classificamos os clientes como fraudadores e bons de acordo com o ponto de corte estimado pela Curva ROC. Obsserva-se que a performance da abordagem bayesiana é mais eficiente nesta amostra ao identificar corretamenta clientes fraudadores e bons.

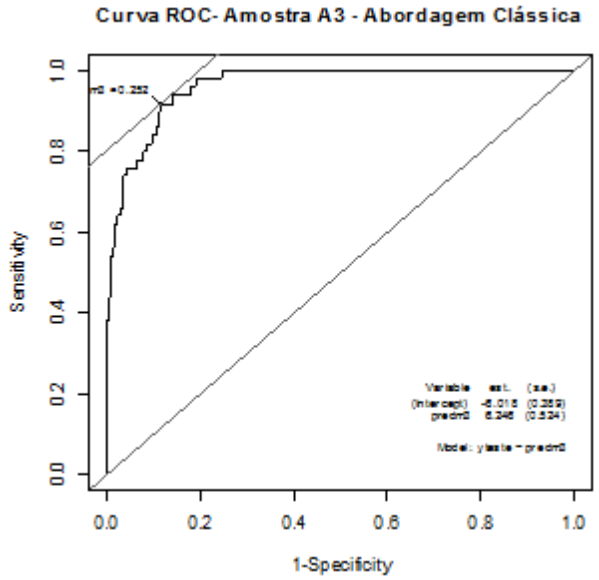
Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	sim	não	Total
	Fraude	44	187	231	Fraude	44	171	215
	não Fraude	6	1813	1819	não Fraude	6	1829	1835
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	88%				88%			
Especificidade	90.6%				91.4%			
Acurácia	90.6%				91.4%			

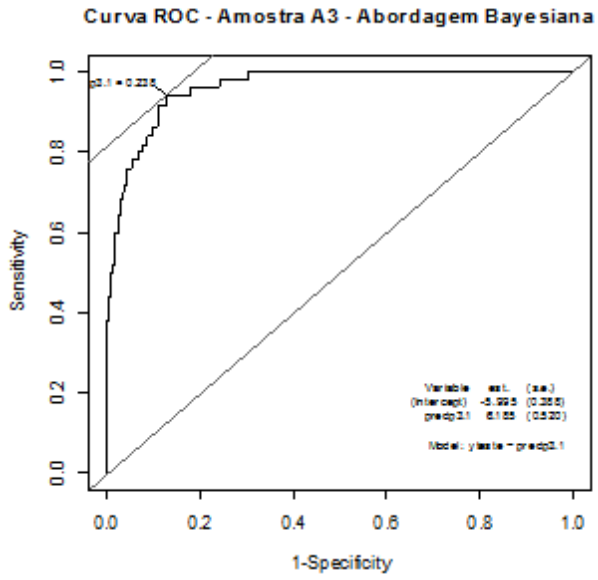
8.2.3 Amostra Balanceada A3 (30/70: 30% *fraudadores* x 70% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 30% de *clientes fraudadores* e 70% de *clientes bons.* O resultado da modelagem clássica e bayesiana está demonstrada na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-9.399358	1.052405	-8.931	< 2e-16	-9.01152	1.047008
x_1	hab_cham	1.166127	0.419186	2.782	0.005404	1.14660	0.423774
x_3	handset	1.330511	0.368417	3.611	0.000305	1.30410	0.375093
x_5	cep	-0.048317	0.005152	-9.379	< 2e-16	-0.04959	0.005129
x_6	ind_marc	3.487677	1.394585	2.501	0.012389	-	-
x_7	ind_prom	6.005836	2.249530	2.670	0.007589	6.53604	2.302735
x_8	ind_faix	3.636171	0.548567	6.628	3.39e-11	3.77058	0.550763
x_9	ind_plan	3.996689	0.863731	4.627	3.71e-06	4.13446	0.881823
x_{10}	ind_sexo	1.899054	0.500278	3.796	0.000147	1.88701	0.510183
x_{11}	ind_cana	2.835177	0.395269	7.173	7.35e-13	2.93502	0.399224
x_{12}	flg_cpf	2.437890	0.598005	4.077	4.57e-05	2.52874	0.614481
x_{13}	flg_cep	3.793875	0.481347	7.882	3.23e-15	3.97333	0.503048
x_{15}	restrica	0.885623	0.248667	3.561	0.000369	0.86107	0.250230
x_{16}	status_c	0.801921	0.132514	6.052	1.43e-09	0.81637	0.132769
x_{17}	credit_s	11.336716	0.850760	13.325	< 2e-16	11.72141	0.866425

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperavámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. No entanto a modelagem bayesiana, apesar de ter produzido um modelo mais econômico, com menos variáveis associadas ao evento, os indicadores de performance do modelo (sensibilidade, acuáricia e especificidade) para esta amostra apresentam valores inferiores ao modelo obtido pela abordagem clássica.





Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem clássica apresenta a especificidade mais alta, com uma melhor classificação de clientes não fraudadores, minimizando o grau de confundimento com os clientes fraudadores.O grau de assertividade do modelo bayesiano não se manteve com da diminuição de casos de fraude utilizados para estimar o modelo.

Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	sim	não	Total
	Fraude	46	233	279	Fraude	46	4	302
	não Fraude	4	1767	1771	não Fraude	4	46	1748
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	92%				92%			
Especificidade	88,3%				87.2%			
Acurácia	88,4%				87.3%			

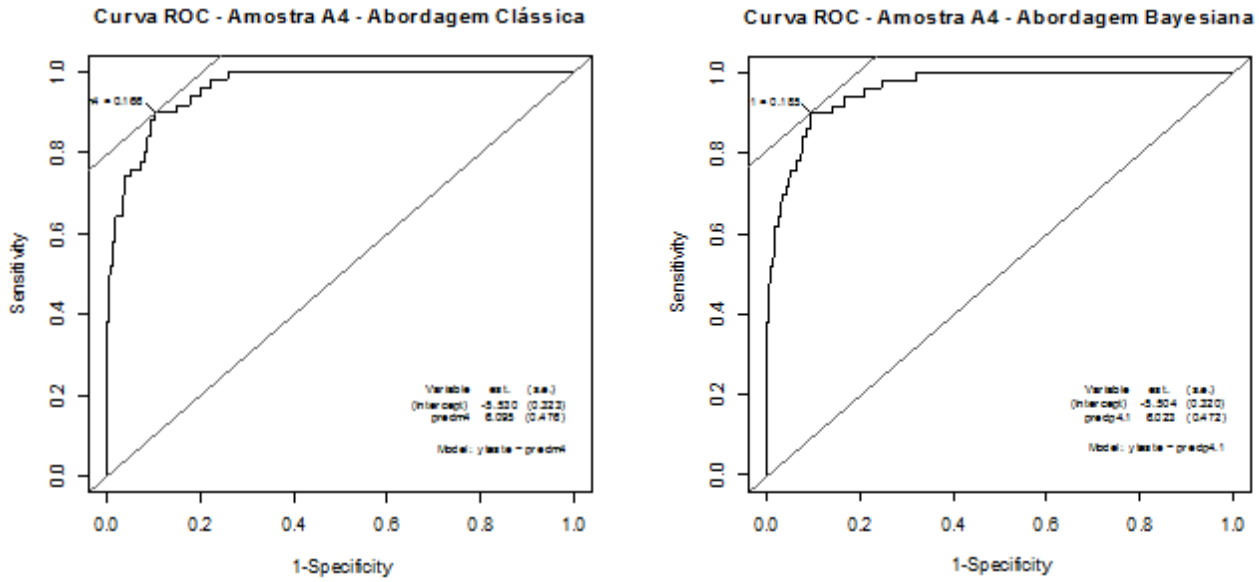
8.2.4 Amostra Balanceada A4 (20/80: 20% *fraudadores* x 80% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 20% de *clientes fraudadores* e 80% de *clientes bons.* O resultado da modelagem clássica e bayesiana está demonstrada na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-12.385699	1.409354	-8.788	< 2e-16	-12.30509	1.432161
x_1	hab_cham	1.683067	0.493053	3.414	0.000641	1.75681	0.511498
x_3	handset	1.544838	0.429208	3.599	0.000319	1.55965	0.432867
x_5	cep	-0.044232	0.006055	-7.305	2.78e-13	-0.04587	0.006146
x_6	ind_marc	3.358376	1.679533	2.000	0.045545	-	-
x_7	ind_prom	8.983129	3.171169	2.833	0.004615	9.70881	3.162058
x_8	ind_faix	4.840864	0.786964	6.151	7.68e-10	5.07212	0.811464
x_9	ind_plan	4.202823	1.035544	4.059	4.94e-05	4.33566	1.050147
x_{10}	ind_sexo	1.648309	0.608423	2.709	0.006746	1.68598	0.612078
x_{11}	ind_cana	2.869368	0.482902	5.942	2.82e-09	2.98825	0.500064
x_{12}	flg_cpf	2.813156	0.634536	4.433	9.28e-06	2.97352	0.644913
x_{13}	flg_cep	4.064408	0.531575	7.646	2.07e-14	4.16972	0.550386
x_{15}	restrica	1.511257	0.293160	5.155	2.54e-07	1.52397	0.294003
x_{16}	status_c	0.757296	0.168034	4.507	6.58e-06	0.78811	0.166851
x_{17}	credit_s	11.946980	1.035000	11.543	< 2e-16	12.56225	1.067216

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperávámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. A modelagem bayesiana produziu um modelo com menos uma variável associada ao evento e, mesmo assim, apresentou um menor índice

de falsos negativos, classificando corretamente 90.6% dos clientes não fraudadores, minimizando o número de casos a serem acompanhados pela área de Análise de Fraude. A capacidade de identificar casos de clientes fraudadores não foi alterada pela exclusão da variável.



Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a especificidade mais alta, com uma melhor classificação de clientes não fraudadores, mimizando o grau de confundimento com os clientes fraudadores.O grau de assertividade do modelo bayesiano se manteve apesar da diminuição de casos de fraude utilizados para estimar o modelo.

Na tabela abaixo classificamos os clientes como fraudadores e bons de acordo com o ponto de corte estimado pela Curva ROC. Obsserva-se que a performance da abordagem bayesiana é mais eficiente nesta amostra ao identificar corretamenta clientes fraudadores e bons.

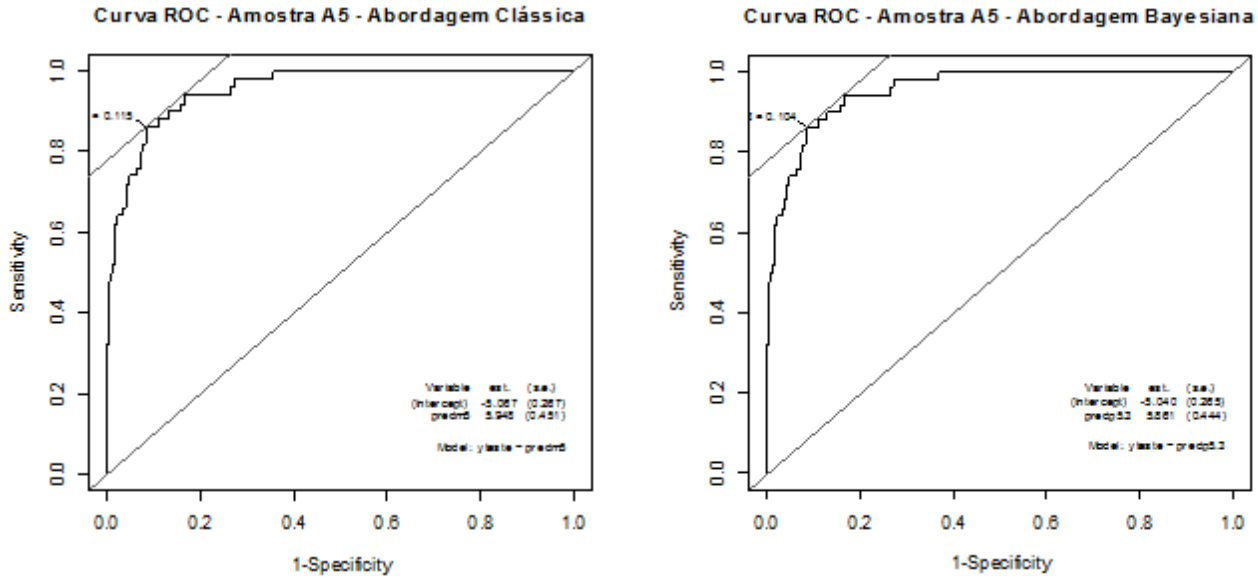
Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	não	sim	Total
	Fraude	44	206	250	Fraude	44	188	232
	não Fraude	6	1794	1800	não Fraude	6	1812	1818
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	88%				88%			
Especificidade	89.7%				90.6%			
Acurácia	89,7%				90.5%			

8.2.5 Amostra Balanceada A5 (10/90: 10% *fraudadores* x 90% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 10% de *clientes fraudadores* e 90% de *clientes bons.* O resultado da modelagem clássica e bayesiana está demonstrada na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-13.819027	1.924830	-7.179	7.00e-13	-14.72998	1.971691
x_1	hab_cham	1.394131	0.769639	1.811	0.070078	1.40115	0.803650
x_3	handset	1.996108	0.547174	3.648	0.000264	2.10306	0.566865
x_5	cep	-0.048566	0.008485	-5.724	1.04e-08	-0.05154	0.008767
x_7	ind_prom	11.796046	4.536229	2.600	0.009311	12.97453	4.692045
x_8	ind_faix	5.399628	1.113442	4.849	1.24e-06	5.73841	1.168156
x_9	ind_plan	4.332861	1.353375	3.202	0.001367	4.57147	1.364169
x_{10}	ind_sexo	2.017907	0.810834	2.489	0.012822	2.1609	0.831416
x_{11}	ind_cana	3.705282	0.652853	5.676	1.38e-08	3.91475	0.668136
x_{12}	flg_cpf	2.786863	0.887838	3.139	0.001696	3.01542	0.913197
x_{13}	flg_cep	4.624078	0.671176	6.890	5.60e-12	4.91338	0.711550
x_{14}	flg_esn	1.811586	0.852341	2.125	0.033551	1.89352	0.873254
x_{15}	restrica	1.842613	0.380650	4.841	1.29e-06	1.94367	0.387736
x_{16}	status_c	1.057900	0.218011	4.853	1.22e-06	1.11200	0.220857
x_{17}	credit_s	12.036165	1.350483	8.912	$<$ 2e-16	12.75784	1.400055

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperavámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. Os modelos obtidos nas aborgens clássica e bayesiana foram muito similares para esta amostra, apresentam as mesmas variáveis com valores muito próximos.



Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a sensibilidade mais alta, com uma melhor classificação de clientes fraudadores, mimizando o grau de confundimento com os clientes fraudadores.O grau de assertividade do modelo bayesiano se manteve apesar da diminuição de casos de fraude utilizados para estimar o modelo.

Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	sim	não	Total
	Fraude	42	167	209	Fraude	43	172	215
	não Fraude	8	1833	1841	não Fraude	7	1828	1835
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	84%				86%			
Especificidade	91,6%				91.4%			
Acurácia	91,5%				91.3%			

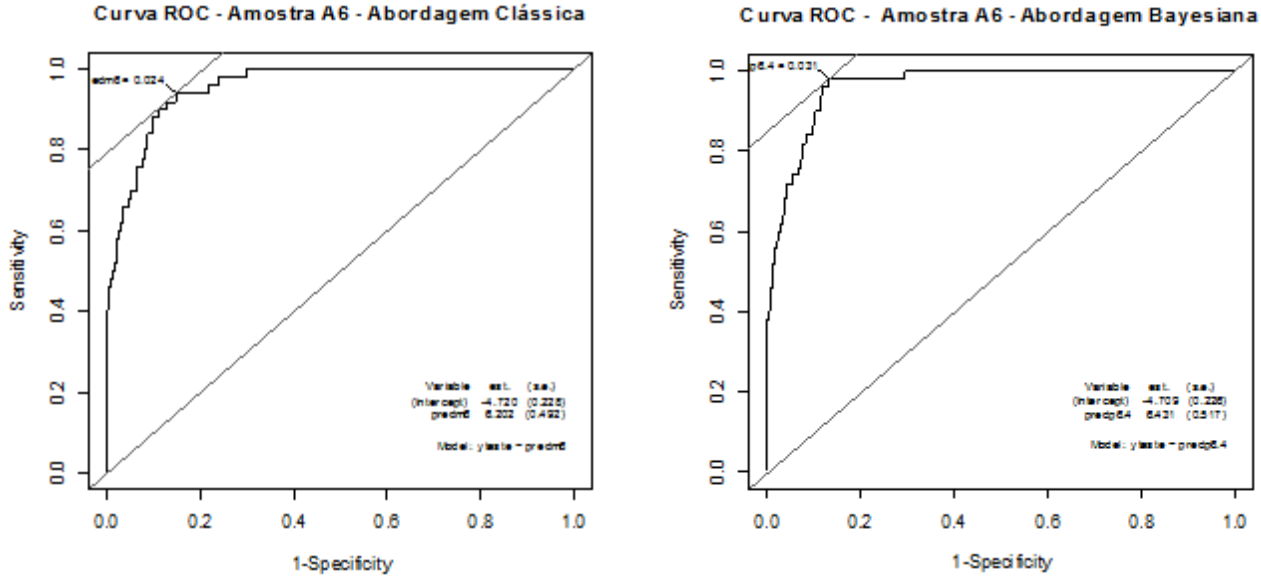
8.2.6 Amostra Balanceada A6 (5/95: 5% *fraudadores* x 95% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 5% de *clientes fraudadores* e 95% de *clientes bons.* O resultado da modelagem clássica e bayesiana está demonstrada na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-11.43164	2.00865	-5.691	1.26e-08	-10.14949	1.755409
x_1	hab_cham	2.83712	0.67779	4.186	2.84e-05	-	-
x_3	handset	1.19880	0.76912	1.559	0.119078	-	-
x_5	cep	-0.04901	0.01066	-4.596	4.31e-06	-0.04406	0.009582
x_7	ind_prom	7.77354	4.87095	1.596	0.110512	-	-
x_8	ind_faix	4.58922	1.39092	3.299	0.000969	4.75275	1.345884
x_9	ind_plan	5.61742	1.64734	3.410	0.000650	5.76452	1.615480
x_{11}	ind_cana	3.08338	0.83424	3.696	0.000219	2.77608	0.810497
x_{12}	flg_cpf	3.20956	0.84727	3.788	0.000152	3.54400	0.859453
x_{13}	flg_cep	4.38393	0.73452	5.968	2.39e-09	4.35867	0.713012
x_{15}	restrica	1.42403	0.48619	2.929	0.003401	1.22817	0.485358
x_{16}	status_c	0.48664	0.29045	1.675	0.093842	0.52349	0.291031
x_{17}	credit_s	11.62069	1.63122	7.124	1.05e-12	11.52454	1.587125

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperávámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. No entanto a modelagem bayesiana produziu um modelo mais econômico, com menos três variáveis associadas ao evento e com maior especificidade e sensibilidade, isto é, classificou mais corretamente os clientes não fraudadores, minimizando o número de casos a serem acompanhados pela área de Análise de

Fraude.



Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a especificidade mais alta, com uma melhor classificação de clientes não fraudadores, mimizando o grau de confundimento com os clientes fruadadores.O grau de assertividade do modelo bayesiano se manteve apesar da diminuição de casos de fraude utilizados para estimar o modelo.

Na tabela abaixo classificamos os clientes como fraudadores e bons de acordo com o ponto de corte estimado pela Curva ROC. Obsserva-se que a performance da abordagem bayesiana é mais eficiente nesta amostra ao identificar corretamenta clientes fraudadores e bons.

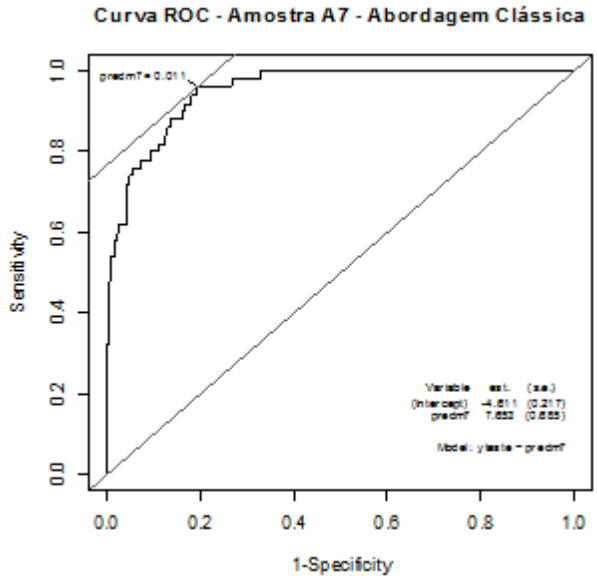
Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	sim	não	Total
	Fraude	47	295	342	Fraude	49	269	318
	não Fraude	3	1705	1708	não Fraude	1	1731	1732
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	94%				98%			
Especificidade	85,3%				86.6%			
Acurácia	85,5%				86.8%			

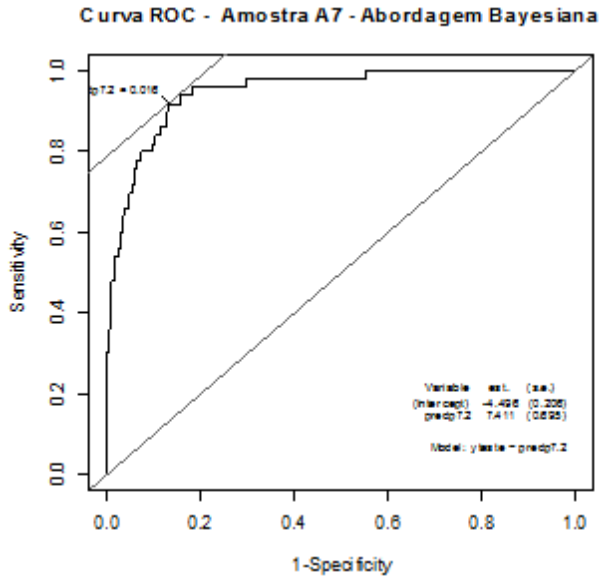
8.2.7 Amostra Balanceada A7 (2.5/97.5: 2.5% *fraudadores* x 97,.5% *bons*)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 2.5% de *clientes fraudadores* e 97.5% de *clientes bons.* O resultado da modelagem clássica e bayesiana está demonstrada na tabela abaixo.

		Modelagem Clássica				Modelagem Bayesiana	
Variável		Parâmetro	σ	z	P(>|z|)	Parâmetro	σ
	Intercepto	-11.39891	2.49836	-4.563	5.05e-06	-9.92736	2.39788
x_1	hab_cham	2.14925	0.85421	2.516	0.011867	1.98824	0.90578
x_5	cep	-0.04347	0.01138	-3.819	0.000134	-0.04774	0.01163
x_6	ind_marc	7.22796	3.22768	2.239	0.025132	-	-
x_8	ind_faix	2.49345	1.32659	1.880	0.060163	3.32871	1.44871
x_9	ind_plan	3.34896	1.99478	1.679	0.093179	-	-
x_{10}	ind_sexo	2.78035	1.36776	2.033	0.042075	3.02211	1.41646
x_{12}	flg_cpf	2.99867	1.05706	2.837	0.004557	3.22306	1.07971
x_{13}	flg_cep	3.57588	0.91298	3.917	8.98e-05	3.72102	0.90161
x_{15}	restrica	1.85575	0.61768	3.004	0.002661	1.94468	0.61786
x_{16}	status_c	1.16377	0.32394	3.593	0.000328	1.24751	0.34372
x_{17}	credit_s	11.27438	2.08831	5.399	6.71e-08	12.02533	2.09251

Na abordagem clássica utilizamos o método stepwise para selecionar as variáveis mais signficativas na classificação dos clientes como fraudadores e não fraudadores, esperávámos que os modelos obtidos pelas duas abordagens fossem semelhantes nas variáveis e parâmetros e poder de classificação. No entanto a modelagem bayesiana produziu um modelo mais econômico, com menos variáveis associadas ao evento e com maior especificidade, isto é, classificou mais corretamente os clinetes não fraudadores, minimizando o número de casos a serem acompanhados pela área de Análise de Fraude.





Apesar dos parâmetros estimados pelas duas abordagens serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a especificidade mais alta, com uma melhor classificação de clientes não fraudadores, mimizando o grau de confundimento com os clientes fruadadores.O grau de assertividade do modelo bayesiano se manteve apesar da diminuição de casos de fraude utilizados para estimar o modelo.

Na tabela abaixo classificamos os clientes como fraudadores e bons de acordo com o ponto de corte estimado pela Curva ROC. Obsserva-se que a performance da abordagem bayesiana é mais eficiente nesta amostra ao identificar corretamenta clientes fraudadores e bons.

Abordagem	Clássica				Bayesiana			
		Fraude				Fraude		
	Modelo	sim	não	Total	Modelo	sim	não	Total
	Fraude	47	379	426	Fraude	45	262	307
	não Fraude	3	1621	1624	não Fraude	5	1738	1743
	Total	50	2000	2050	Total	50	2000	2050
Sensibilidade	94%				90%			
Especificidade	81%				86.9%			
Acurácia	81.4%				87%			

8.2.8 Considerações

Como as prioris utilizadas foram não informativas, não era esperado encontrar discrepâncias entre os valores dos modelos ajustados por inferência clássica e bayesiana, como foi verificado neste estudo.

No entando à medida que o evento de interesse tem uma baixa incidência na amostra de treinamento a abordagem bayesiana tende a ter melhores resultados que a abordagem clássica, isto é, a especificidade do modelo é mantida apesar da pouca incidência de casos de interesse, classificando corretamente em clientes com baixa propensão a fraude como bons clientes.

8.3 Segunda fase: Análise Bayesiana considerando Priori Informativa

Da perpectiva bayesiana, os dados históricos podem ser muito úteis em interpretar os resultados atuais do evento de interesse. A utilização de priori informativas para dados binários correlacionados a resposta, onde a especificação é baseada na existência de um estudo anterior similar (com as covariadas e respostas) ao estudo atual (Dados históricos).

8.3.1 Motivação Priori Informativas

A grande motivação para utilizar a inferência bayesiana na determinação do comportamento do fraudador refere-se a volatilidade de seu modo de operação, o comportamento da fraude é muito dinâmico, modificando-se num curto espaço de tempo. Um outro problema a ser enfrentado refere-se a disponibilidade ds informações pertinentes, seja por histórico insuficiente ou incidência de casos muito baixa.

A abordagem bayesiana possibilita construir o conhecimento sobre o evento Fraude a partir da inclusão de informações históricas incrementais em funções encadeadas. Partindo-se de uma modelagem com prioris não informativas obtemos a função de distribuição a posteriori, que será utilizada como informação inicial do evento, como um conhecimento a priori. Desta forma, com carga incremental de dados controi-se um método eficaz para inferir e internalizar novos padrões de fraude. Na figura 8.2 observa-se o conceito de construção do conhecimento sobre o evento de interesse (fraude de subscrição) utilizando-se da abordagem bayesiana.

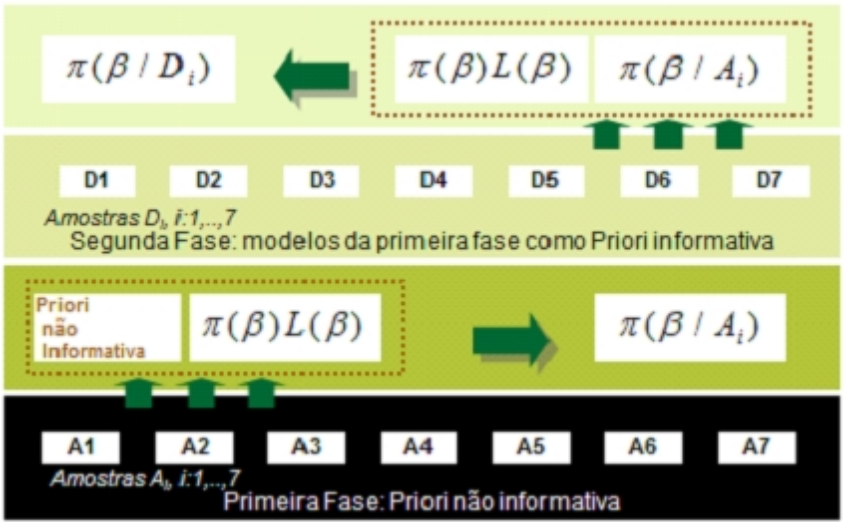


Figura 8.2: Modelos Encadeados

8.3.2 Resultados

Para implementar a utilização de uma priori informativa, a partir de dados históricos subdividimos a aplicação em duas fases. Na primeira fase foram estimados os hiper-

parâmetros de um modelo de regressão logística partindo-se de prioris não informativas. O resultado das estimativas dos hiperparâmetros foram utilizadas como prioris informativas nesta fase da aplicação, de forma que cada amostra D_j gerou i modelos estimados baseado nos resultados obtidos na primeira fase. Desta forma, cada amostra D_j gerou i modelos de regressão logística de acordo com o esquema exemplificado na figura 8.3:

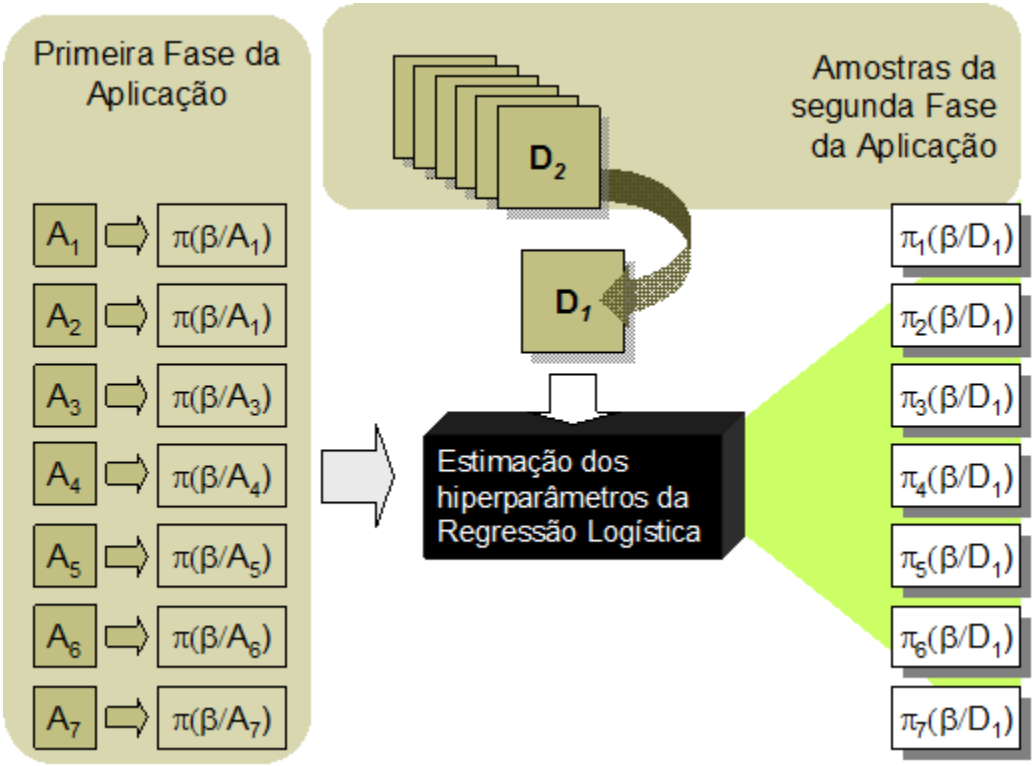


Figura 8.3: Visão da Aplicação de Inferência Bayesiana

As amostras extraídas aleatoriamente nesta segunda fase da aplicação constituem-se em um subconjunto da base de treinamento do modelo original, não utilizada na primeira fase deste trabalho.Vale ressaltar que a base treinamento original, utilizada para estimar os parâmetros do modelo com melhor performance, foi estruturada de forma a apresentar um *oversampling* 30/70, isto é, amostra balanceada com a incidência de 30% do total de casos como clientes *fraudadores.* Nesta segunda fase do trabalho utilizamos amostras balanceadas 50/50, isto é com 50% de incidência de *clientes fraudadores.* Nesta fase o objetivo é desenvolver um modelo capaz de identificar padrões / comportamento do

cliente fraudador com um volume de casos para estudo reduzido. Desta forma, procuramos reproduzir a situação real de uma empresa de Telecomunicações preocupada em identificar o mais rapidamente novos padrões / comportamentos em eventos raros de fraude

Assim a utilização de amostras balanceadas e pequenas visa proporcionar alternativas para tratar a volatidade do evento de interesse, sendo capaz de classificar corretamente os padrões de fraude em constante mudança e reciclagem.

Para esta fase do trabalho utilizou-se a mesma estrutura do modelo estimado na primeira fase. Foi utilizada a técnica de *oversampling* para garantir a capacidade preditiva do modelo estimado, utilizando amostras balanceadas artificialmente. A tabela a seguir descreve as amostras utilizadas.

Amostras	Casos de Fraude	Casos de não Fraude	Total de Casos	Incidência de Fraude
D_1	1238	1238	2476	50%
D_2	884	866	1750	50.5%
D_3	609	607	1216	50.1%
D_4	376	370	746	50.4%
D_5	129	127	256	50.4%
D_6	65	58	123	52.8%
D_7	31	31	62	50%

O valor estimado dos hiperpâmetros iniciais foram obtidos utilizando o algoritmo de Metropolis-Hastings, com prioris não informativas, conforme descrito em "Primeira fase: Comparação da Abordagem Clássica e Bayesiana".

8.3.3 Amostra Balanceada D_1 (50/50, n: 2476 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamnaho da amostra é 2476 clientes, sendo 1278 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira

fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo.

Modelo	Priori	Sensibilidade	Especificidade	Acurácia	KM	Ponto de corte
$\pi_1(\beta/D_1)$	$\pi(\beta/A_1)$	96%	85.6%	85.8%	0.963	0.331
$\pi_2(\beta/D_1)$	$\pi(\beta/A_2)$	94%	87.3%	87.4%	0.963	0.371
$\pi_3(\beta/D_1)$	$\pi(\beta/A_3)$	94%	87.6%	87.8%	0.964	0.390
$\pi_4(\beta/D_1)$	$\pi(\beta/A_4)$	94%	86%	86.1%	0.962	0.341
$\pi_5(\beta/D_1)$	$\pi(\beta/A_5)$	92%	84.7%	84.9%	0.959	0.293
$\pi_6(\beta/D_1)$	$\pi(\beta/A_6)$	98%	86%	86.2%	0.963	0.369
$\pi_7(\beta/D_1)$	$\pi(\beta/A_7)$	94%	86.9%	87%	0.959	0.425

8.3.4 Amostra Balanceada D_2 (50/50, n: 1750 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamanho da amostra é 1750 clientes, sendo 884 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo.

Modelo	Priori	Sensibilidade	Especificidade	Acurácia	KM	Ponto de corte
$\pi_1(\beta/D_2)$	$\pi(\beta/A_1)$	94%	88%	88.2%	0.963	0.409
$\pi_2(\beta/D_2)$	$\pi(\beta/A_2)$	94%	87.5%	87.7%	0.962	0.395
$\pi_3(\beta/D_2)$	$\pi(\beta/A_3)$	94%	88.4%	88.5%	0.964	0.423
$\pi_4(\beta/D_2)$	$\pi(\beta/A_4)$	94%	86.8%	86.9%	0.962	0.364
$\pi_5(\beta/D_2)$	$\pi(\beta/A_5)$	94%	84.5%	84.8%	0.959	0.295
$\pi_6(\beta/D_2)$	$\pi(\beta/A_6)$	96%	87%	87.2%	0.963	0.384
$\pi_7(\beta/D_2)$	$\pi(\beta/A_7)$	94%	86.8%	87%	0.959	0.430

8.3.5 Amostra Balanceada D_3 (50/50, n: 1216 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamnaho da amostra é 1216 clientes, sendo 609 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo.

Modelo	Priori	Sensibilidade	Especificidade	Acurácia	KM	Ponto de corte
$\pi_1(\beta/D_3)$	$\pi(\beta/A_1)$	94%	87.3%	87.5%	0.962	0.385
$\pi_2(\beta/D_3)$	$\pi(\beta/A_2)$	96%	84.5%	84.7%	0.962	0.291
$\pi_3(\beta/D_3)$	$\pi(\beta/A_3)$	94%	87.7%	88%	0.963	0.411
$\pi_4(\beta/D_3)$	$\pi(\beta/A_4)$	94%	85.4%	85.6%	0.961	0.322
$\pi_5(\beta/D_3)$	$\pi(\beta/A_5)$	86%	91.3%	91.1%	0.958	0.518
$\pi_6(\beta/D_3)$	$\pi(\beta/A_6)$	98%	85.5%	85.8%	0.962	0.352
$\pi_7(\beta/D_3)$	$\pi(\beta/A_7)$	92%	88.8%	88.9%	0.958	0.490

8.3.6 Amostra Balanceada D_4 (50/50, n: 746 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamnaho da amostra é 746 clientes, sendo 376 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os

resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo.

Modelo	Priori	Sensibilidade	Especificidade	Acurácia	KM	Ponto de corte
$\pi_1(\beta/D_4)$	$\pi(\beta/A_1)$	92%	89.7%	89.7%	0.962	0.464
$\pi_2(\beta/D_4)$	$\pi(\beta/A_2)$	92%	89%	89.1%	0.962	0.439
$\pi_3(\beta/D_4)$	$\pi(\beta/A_3)$	92%	89.7%	89.7%	0.963	0.466
$\pi_4(\beta/D_4)$	$\pi(\beta/A_4)$	94%	85.3%	85.5%	0.961	0.323
$\pi_5(\beta/D_4)$	$\pi(\beta/A_5)$	94%	84.3%	84.5%	0.959	0.289
$\pi_6(\beta/D_4)$	$\pi(\beta/A_6)$	98%	86.3%	86.5%	0.961	0.374
$\pi_7(\beta/D_4)$	$\pi(\beta/A_7)$	92%	89.1%	89.2%	0.958	0.491

8.3.7 Amostra Balanceada $D5$ (50/50, n: 256 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamnaho da amostra é 256 clientes, sendo 129 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo.

Modelo	Priori	Sensibilidade	Especificidade	Acurácia	KM	Ponto de corte
$\pi_1(\beta/D_5)$	$\pi(\beta/A_1)$	92%	88.7%	88.7%	0.96	0.482
$\pi_2(\beta/D_5)$	$\pi(\beta/A_2)$	92%	87.6%	87.7%	0.959	0.439
$\pi_3(\beta/D_5)$	$\pi(\beta/A_3)$	92%	89.1%	89.2%	0.962	0.489
$\pi_4(\beta/D_5)$	$\pi(\beta/A_4)$	94%	87.5%	87.6%	0.96	0.400
$\pi_5(\beta/D_5)$	$\pi(\beta/A_5)$	94%	84.8%	85%	0.957	0.318
$\pi_6(\beta/D_5)$	$\pi(\beta/A_6)$	94%	89.1%	89.2%	0.96	0.491
$\pi_7(\beta/D_5)$	$\pi(\beta/A_7)$	90%	88.8%	88.7%	0.955	0.571

8.3.8 Amostra Balanceada D_6 (50/50, n: 123 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamnaho da amostra é 123 clientes, sendo 65 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo.

Modelo	Priori	Sensibilidade	Especificidade	Acurácia	KM	Ponto de corte
$\pi_1(\beta/D_6)$	$\pi(\beta/A_1)$	96%	84.3%	84.5%	0.958	0.228
$\pi_2(\beta/D_6)$	$\pi(\beta/A_2)$	96%	83.1%	83.4%	0.958	0.188
$\pi_3(\beta/D_6)$	$\pi(\beta/A_3)$	96%	84.2%	84.4%	0.959	0.225
$\pi_4(\beta/D_6)$	$\pi(\beta/A_4)$	92%	83.7%	83.9%	0.956	0.202
$\pi_5(\beta/D_6)$	$\pi(\beta/A_5)$	84%	90.1%	90.5%	0.951	0.397
$\pi_6(\beta/D_6)$	$\pi(\beta/A_6)$	96%	83.4%	83.7%	0.958	0.224
$\pi_7(\beta/D_6)$	$\pi(\beta/A_7)$	88%	88%	88%	0.949	0.436

8.3.9 Amostra Balanceada $D7$ (50/50, n: 62 clientes)

Esta amostra foi obtida aleatoriamente da base de treinamento da Modelagem Original, através da metodologia de *oversampling*, originando uma amostra de treinamento balanceada com 50% de *clientes fraudadores* e 50% de *clientes bons.* Vale ressaltar que o tamnaho da amostra é 62 clientes, sendo 31 classificados como *clientes fraudadores.* A estimativa dos hiperpâmetros, resultado da modelagem bayesiana efetuada na primeira fase deste trabalho, foram utilizados como prioris informativas neste segundo estágio. Os resultados obtidos para esta primeira amostra estãodemonstrados na tabela abaixo

Modelo	Priori	Sensibilidade	Especificidade	KM	Ponto de corte
$\pi_1(\beta/D_6)$	$\pi(\beta/A_1)$	96%	80.7%	0.958	0.000
$\pi_2(\beta/D_6)$	$\pi(\beta/A_2)$	96%	80.7%	0.86	0.000
$\pi_3(\beta/D_6)$	$\pi(\beta/A_3)$	92%	83.7%	0.859	0.921
$\pi_4(\beta/D_6)$	$\pi(\beta/A_4)$	92%	83.7%	0.859	0.921
$\pi_5(\beta/D_6)$	$\pi(\beta/A_5)$	96%	80.7%	0.86	0.000
$\pi_6(\beta/D_6)$	$\pi(\beta/A_6)$	92%	85.3%	0.879	0.000
$\pi_7(\beta/D_6)$	$\pi(\beta/A_7)$	92%	81.3%	0.926	0.334

Observa-se que nesta amostra o resultado da modelagem tem pior performance que os modelos estimados a partir de amostras maiores. O ponte de corte ora se apresenta próximo a zero ora se apresenta próximo a 1. Apenas um modelo apresentou estatísticas compatíveis com os modelos estimados das amostras anteriores. Nesta situação não foi possível calcular a acurácia dos modelos.

Capítulo 9

Considerações finais

Como o modelo ajustado por inferência clássica tem uma porcentagem de acerto superior a 95% não foi possível aprimorar o desempenho da predição.

Na primeira fase, apesar dos parâmetros estimados pelas abordagens classica e bayesiana serem muito próximos a performance do modelo obtido através da modelagem bayesiana apresenta a especificidade mais alta, com uma melhor classificação de clientes não fraudadores, mimizando o grau de confundimento com os clientes fruadadores.O grau de assertividade do modelo bayesiano se manteve apesar da diminuição de casos de fraude utilizados para estimar o modelo.

O principal objetivo de comparar a performance da modelagem baseada em inferência bayesiana é extrair informações / padrões de comportamento dos eventos ditos raros, com uma incidência muito baixa na população, mas, que no caso de fraude, ocasionam perdas financeiras consideráveis para as empresas de Telecomunicações Celures.

Desta forma a pouca representatividade de casos de fraude é uma situação comum numa empresa de telecomunicações, e a espera para completar um número de safras superior a 3 meses pode gerar perdas financeiras significativas e misturar padrões de comportamentos distintos, uma vez que o fenômeno "fraude"está em constante mudança e reciclagem. O objetivo deste estudo é proporcionar alternativas para tratar a volatidade do evento de interesse.

A utilização de priori informatiava, e cargas incrementais de dados mostrou ser um excelente método de inferir e internalizar novos padrões de fraude. Mesmo em amostras pequenas, inferior a 300 clientes observou-se que a abordagem bayesiana produz esti-

madores consistentes e eficientes para discriminar os clientes bons dos fraudadores.

Mesmo em situações em que a base histórica de eventos de fraude é superior a 12 meses, vale lembrar que o comportamento do fraudador está sempre se modificando, visando obter o máximo de benefícios de seu relacionamento com a Operadora de Telecomunicações. Assim sendo, iinformações de comportamentos mais antigos podem mascarar novos padrões de comportamento. A grande motivação para utilizar a inferência bayesiana na determinação do comportamento do fraudador refere-se a volatilidade de seu modo de operação, o comportamento da fraude é muito dinâmico, modificando-se num curto espaço de tempo. Um outro problema a ser enfrentado refere-se a disponibilidade ds informações pertinentes, seja por histórico insuficiente ou incidência de casos muito baixa.

A abordagem bayesiana permite construir o conhecimento sobre o evento Fraude com a inclusão de informações históricas incrementais em funções encadeadas. Partindo-se de uma modelagem com prioris não informativas para estimar a função de distribuição a posteriori e utilizando esta informação iinicial do evento como um conhecimento a priori, com carga incremental de dados controi-se um método eficaz para inferir e internalizar novos padrões de fraude.

Referências Bibliográficas

[1] G. Piatetsky-Shapiro. Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. AI Magazine, Vol. 11, No. 5, Jan. 1991, Special issue, 68-70

[2] S. Kelly. Data Warehouse applications in the telecommunications industry. Proc. Conf. Commercial Parallel Processing. London, IBC, 1995.

[3] R. Srikant and R. Agrawal. Mining generalized association rules. Proc. 21st Very Large Databases (VLDB) Conf., Zurich, Switzerland, 1995.

[4] D. Michie, D. J. Spiegelhalter and C. C. Taylor. Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood, 1994.

[5] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering . Machine Learning, 2, 1987, 139-172.

[6] Fayyad, U. M., Piatetsky Shapiro, G., Smyth, P. & Uthurusamy, R. – "Advances in Knowledge Discovery and Data Mining", AAAIPress, The Mit Press, 1996.

[7] P. Langley. Elements of Machine Learning. Morgan Kaufmann, 1996.

[8] J. W. Shavlik and T. G. Diettrich. (Eds.) Readings in Machine Learning. San Mateo, CA:Morgan Kaufmann, 1990.

[9] J. F. Elder IV and D. Pregibon. A statistical perspective on knowledge discovery in data bases.In: U. M. Fayyad et al. (Ed.) Advances in Knowledge Discovery and Data Mining, 83-113. AAAI/MIT Press, 1996.

[10] H-Y. Lee, H-L. Ong and L-H. Quek. Exploiting visualization in knowledge discovery. Proc. 1st Int. Conf. Knowledge Discovery and Data Mining (KDD-95), 198-203. AAAI, 1995.

[11] Haykin, S., Neural Networks: A Comprehensive Foundation, Macmillan College PublishingCompany, New York, NY, 1994.

[12] D. Rumelhart and McClelland. (Eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press, 1986.

[13] N. J. Nilsson. Principles of Artificial Intelligence. Palo Alto, CA: Tioga, 1980.

[14] D. E. Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[15] Ming-Hui Chen and Dipak K. Dey, "Variable Selection for Multivariate Logistic Regression Models" in Journal of Statistical Planning and Inference, 111, 37-55, 2003

[16] Joseph G. Ibrahim and Ming-Hui Chen, "Power Prior Distributions for Regression Models in Statistical Science 2000, Vol. 15, No. 1, 46–60

[17] RJ Bolton and DJ Hand. Statistical fraud detection: A review. Statistical Science, 17(3):235–255, 2002. 331

[18] Swets, J. A. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. New Jersey: LEA, 1996.

[19] Metz, C. E. "Statistical Analysis of ROC Data in Evaluating Diagnostic Performance." Multiple Regression Analysis: Applications in the Health Sciences, number 13, edited by Donald E. Herbert and Raymond H. Myers. 365–384. American Institute of Physics, 1986

[20] Metz, C. E. "Basic Principles of ROC Analysis," Seminars in Nuclear Medicine, VIII (4):283–298 (1978).

[21] Green, D. M. and J. A. Swets. "Signal Detection Theory and Psychophysics". New York: Robert E. Krieger Publishing Company, 1973.

[22] Neyman, J. and Pearson, E. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses". Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231: 289–337

[23] Egan, James P. "Signal Detection Theory and ROC Analysis". New York: Academic Press, 1975.

[24] Paulino, C. D. , Turkamn, A. A. e Murteira, B. (2003) Estatística bayesiana. Fundação Calouste Gulbenkian. Lisboa, 446p.

[25] Gilks, W. R. 1996. "Full conditional distributions." In Markov chain Monte Carlo in practice, ed.W. R. Gilks, S. Richardson and D. J. Spiegelhalter. London: Chapman & Hall pp. 75–88.

[26] Bigus, J. P., Data Mining with Neural Network – Solving Business Problems from Application Development to Decision Support, McGraw-Hill, 1996

[27] Holland, J. H., Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA: 1992.

[28] V. Dhar, R. Stein, Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice-Hall, 1997.

[29] Perelmuter, G., Redes Neurais Aplicadas ao Reconhecimento de Imagens Bidimensionais. Dissertação de Mestrado, DEE, PUC – Rio, 1996.

[30] Dayhoff, J., Neural Network Architectures: Na Introduction, Van Nostrand Reinhold, New York, NY: 1990.

[31] L. Fu. Neural Networks in Computer Intelligence. MacGraw-Hill, 1994.

[32] Haykin, S., Neural Networks: A Comprehensive Foundation, Macmillan College Publishing Company, New York, NY, 1994.

[33] V. Dhar, R. Stein, Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice-Hall, 1997

[34] Freeman, J. A., and D. M. Skapura, Neural Networks: Algorithms, Applications, and Programming Techniques, Addison-Wesley, Reading, MA: 1992.

[35] V. Tresp, J. Hollatz, and S. Ahmad. Representing Probabilistic Rules with Networks of Gaussian Basis Functions. Machine Learning, 27, Page 173, 1997. Kluwer Academic Publishers, Boston.

[36] G. G. Towell, and J. W. Shavlik. Extracting Refined Rules from Knowledge-Based Neural Networks. Machine Learning, 13, Page 71, 1993. Kluwer Academic Publishers, Boston.

[37] Brazdil, P. B. Construção de modelos de decisão a partir de dados. Disponível em: <http://www.niaad.liacc.up.pt/~pbrazdil/Ensino/ML/ModDecis.html. Acesso em junho 2003.

[38] Garcia, S. C.; Alvares, L. O. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. Disponível em: <http://www.inf.ufrgs.br/pos/SemanaAcademica/Semana2000/SimoneGarcia/>. Acesso em junho 2003

[39] Ingargiola, G. Building classification models: ID3 and C4.5. Disponível em: <http://www.cis.temple.edu/~ingargio/cis587/readings/id3- c45.html>. Acesso em julho 2003.

[40] Breiman, L.; Friedman, J.H.; Olshen, R. A. Classification and Regression Trees. Belmont, CA: Wadsworth, 1984, 358p.

[41] McLachlan, G. Discriminant Analysis and Statistical Pattern Recognition. New York: John Wiley & Sons, 1992, 526p.