Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Reamostragem e Imputação de Dados em Caso de Eventos Raros

por

Camyla Serpa Oliveira

Brasília

2013

Camyla Serpa Oliveira

Reamostragem e Imputação de Dados em Caso de Eventos Raros

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2013

Dedico esse trabalho à minha mãe, por ter se esforçado diariamente a me ensinar o valor dos estudos.
"O conhecimento é a única coisa que não se perde, ninguém rouba, ninguém toma. Um dia conquistado, será sempre seu!"

Camyla Serpa Oliveira

Agradecimentos

Registro aqui minha eterna gratidão aos meus pais, por terem segurado minha mão e me levado ao primeiro dia de aula, desde então sinto vocês ao meu lado e foi esse carinho e amor que me trouxe até aqui! Agradeço também aos meus tios, o apoio de vocês foi essencial para cada conquista da minha vida.

Sou muito grata à minha vida, por tantas oportunidades e alegrias ter me proporcionado! Pelos amigos colocados em meu caminho que fizeram de toda essa trajetória um livro de histórias.

Meus sinceros agradecimentos ao Prof. Dr. Alan Ricardo da Silva, pela paciência e atenção oferecidas à mim em cada semana de trabalho.

Termino com um agradecimento especial à ESTAT Consultoria Junior em Estatística, que me construiu enquanto ser humano! Espero daqui para frente encontrar experiências tão enriquecedoras na minha vida quanto foi a ESTAT nestes últimos anos.

Resumo

Um problema que frequentemente dificulta a análise de concessão de crédito é o desbalanceamento presente em base de dados bancários, isto acontece devido à baixa ocorrência de clientes inadimplentes nas carteiras das instituições financeiras. Por mais que essa realidade seja essencial para a saúde financeira da instituição, os modelos estatísticos utilizados nas análises desses dados perdem poder de predição, tornando-se difícil construir modelos de probabilidade para esses indivíduos e, com isso, há dificuldades em avaliar estes "maus"clientes. A fim de diagnosticar o desempenho dos modelos de risco nessas carteiras que possuem pouca inadimplência, foi proposto utilizar as técnicas de Reamostragem e de imputação de observações sintéticas SMOTE (*Synthetic Minority Over-sampling Technique*).

Com a aplicação das metodologias propostas, a técnica de Reamostragem se mostrou mais adequada no tratamento de bases de dados desbalanceadas, por produzir percentuais de acertos tão bons quantos a técnica SMOTE mas sem suas limitações. Após utilizada a técnica de Reamostragem houve uma melhora no desempenho do modelo, de tal forma que tornou-se viável a análise de eventos raros. O modelo que antes possuía uma boa acurácia apenas para a classe de frequência predominante, tournou-se um bom preditor também para a classe de baixa frequência.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

O crédito, administrado por Bancos e demais Instituições Financeiras, tem importante papel no processo de manutenção da economia de um país, sendo ele o combustível para estimular o consumo das pessoas, o nível de produção das empresas e, por consequência, o aquecimento da economia. Sua função essencial é promover a otimização dos capitais existentes, melhor alocando recursos àquelas atividades que mais necessitam.

Para que as entidades de proteção ao crédito cumpram sua função social, não basta simplesmente armazenar dados, elas devem analisá-los e agregar valor aos mesmos, oferecendo ao concedente de crédito soluções que viabilizem o crédito ao maior número de consumidores e reduzindo os custos da inadimplência.

Um problema, que frequentemente dificulta a análise das concessões de crédito, é o desbalanceamento presente em base de dados bancárias devido a ocorrência de eventos raros, sendo que a base será considerada desbalanceada se possuir classes que não são igualmente representadas, contendo uma ou várias classes com quantidade inferior às demais. É comum que as carteiras das instituições tenham poucos *defaults*, isto é, clientes que atrasaram o pagamento do empréstimo por mais de 60 dias.

Apesar de essa ser uma realidade desejada para a saúde financeira da instituição, torna-se difícil construir um modelo de probabilidade para esses indivíduos e, com isso, há dificuldades em avaliar estes "maus" clientes.

Ao dispor de informações fidedignas, processadas e disponibilizadas de maneira segura, o concedente de crédito pode melhor quantificar os riscos e assim reduzir os custos decorrentes da inadimplência. O tomador, por sua vez, beneficia-se pelo fato de poder ter sua capacidade de pagamento adequadamente avaliada e consequentemente obter condições de prazo e juros que melhor se adequam a sua realidade. Para isso, os operadores de crédito estão sempre buscando aprimorar suas avaliações de crédito.

A fim de diagnosticar o desempenho dos modelos de risco em carteiras que possuem pouca inadimplência, propõe-se utilizar as técnicas de reamostragem e de imputação de observações sintéticas SMOTE (*Synthetic Minority Over-sampling Technique*).

1.1 OBJETIVOS

O objetivo geral deste trabalho é aplicar métodos de amostragem que viabilizem a análise de eventos raros em modelos logísticos através de simulações.

Os objetivos específicos são:

- Utilizar a técnica de imputação de dados *SMOTE*;
- Utilizar a técnica de reamostragem;
- Comparar as duas técnicas;

- Realizar as análises utilizando o *software* SAS 9.2.

Capítulo 2

REGRESSÃO LOGÍSTICA

2.1 INTRODUÇÃO

Como apontado por King and Zeng (2001), embora as propriedades estatísticas dos modelos de regressão linear sejam invariantes à média da variável dependente, o mesmo não é verdade para os modelos de variáveis dependentes binárias. A média da variável binária é a frequência relativa dos eventos, e por isso uma base de dados desbalanceada traz consequências importantes para as análises produzidas.

A Regressão Logística é uma técnica que produz, a partir de um conjunto de observações estudadas, um modelo que permite a predição de valores de uma variável Y a partir de uma ou mais variáveis. Uma primeira abordagem será feita para o caso de regressão logística simples, onde há apenas uma variável explicativa. Este tipo de regressão se diferencia da regressão linear pois a variável resposta trabalhada é categórica.

A variável categórica é assim definida por poder ser mensurada usando um número limitado de categorias, no presente estudo a váriável categórica Y é classi-

ficada como *dummy* pois possui apenas duas opções de eventos:

$$\begin{cases} Y_i = 0 \; cliente \; adimplente \\ Y_i = 1 \; cliente \; inadimplente \end{cases}$$

2.2 O MODELO DE REGRESSÃO LOGÍSTICA

O modelo de regressão logística é utilizado para estimar a probabilidade dos eventos dicotômicos ocorrerem, onde $Y_i \in \{0, 1\}$ e $x \in \Re$. Sendo Y_i a variável resposta (dependente), e X_i a variável explicativa (independente), o modelo linear que assume $E(\epsilon_i) = 0$ será descrito por:

$$E(Y_i) = \beta_o + \beta_1 X_i, \tag{2.1}$$

onde cada Y_i tem distribuição $Bernoulli(1, \pi)$ com probabilidade de sucesso $P(Y_i = 1) = \pi_i$ e probabilidade de fracasso $P(Y_i = 0) = 1 - \pi_i$. O interesse está centrado em verificar o valor esperado de Y, sendo assim calcula-se a esperança

$$E(Y_i) = \pi_i \tag{2.2}$$

e igualando (2.1) a (2.2):

$$E(Y_i) = \pi_i = \beta_0 + \beta_1 X_i. \tag{2.3}$$

A função resposta é denominada função logística, cuja a expressão é

$$\pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 + x_i}}. \tag{2.4}$$

Uma propriedade importante é que a função logística pode ser linearizada, denotando-se $E(y)$ por π, pois a resposta média é a probabilidade quando a variável resposta em questão é binária. A transformação:

$$g(x) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i \tag{2.5}$$

é denominada transformação logit da probabilidade π e gera a função $g(x)$ que possui muitas propriedades importantes de um modelo de regressão linear.

Partindo de $0 < P(Y_i = y_i|x) < 1$, pode-se verificar que a função logaritmo é estritamente crescente. Sendo assim o passo seguinte é estimar os estimadores de máxima verossimilhança de β_0 e β_1 que maximizem o logaritmo da função de máxima verossimilhança. Utilizando $x = x_i$ tem-se que $\pi(x)$ definido em (2.4) fornece as probabilidades

$$\pi(x) = P(Y_i = 1|x)$$

$$1 - \pi(x) = P(Y = 0|x).$$

Com isso define-se a função de verossimilhança por

$$l(\beta) = \prod_i \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{2.6}$$

aplicando o logaritmo

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n}\{y_i ln[\pi(x_i)] + (1 - y_i)ln[1 - \pi(x_i)]\}. \tag{2.7}$$

Para maximizar a função de máxima verossimilhança deriva-se em relação aos parâmetros do modelo e iguala-se as expressões a zero, como feito em (Hosmer and Lemeshow, 2000)

$$\sum [y_i - \pi(x_i)] = 0$$

e

$$\sum x_i[y_i - \pi(x_i)] = 0.$$

Uma importante consequência dessas equações é que

$$\sum y_i = \sum \pi(x_i). \tag{2.8}$$

No entanto essas expressões são não-lineares nos parâmetros, e para resolvê-las é preciso recorrer a métodos numéricos.

2.3 ODDS RATIO

Encontradas as estimativas, substitui-se esse valores em (2.4) para encontrar os valores ajustados. A função de resposta ajustada é dado por:

$$\hat{\pi} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1+x_i}}. \tag{2.9}$$

Usando a transformação logit em (2.5), a função resposta é ajustada por:

$$\hat{\pi} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{2.10}$$

sendo

$$\hat{\pi} = log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right). \tag{2.11}$$

Este valor representa a estimativa da probabilidade de sucesso no evento.

Considerando o valor da função resposta ajustada (2.10), assumindo $X = X_j$

$$\hat{\pi}(X_j) = \hat{\beta}_0 + \hat{\beta}_1 X_j \tag{2.12}$$

e $X = X_j + 1$

$$\hat{\pi}(X_j + 1) = \hat{\beta}_0 + \hat{\beta}_1(X_j + 1). \tag{2.13}$$

De acordo com (2.12), o logaritmo da chance (*odds*) estimada quando $X = X_j$ é chamado de $log(\text{chance}_1)$, e seguindo a mesma linha de raciocínio, (2.13) é o logaritmo da chance estimada quando $X = X_j$ chamado de $log(\text{chance}_2)$. A diferença entre esses dois valores ajustado é dado por:

$$log(chance_2) - log(chance_1) = log\left(\frac{chance_2}{chance_1}\right) = \hat{\beta}_1.$$

Aplicando o anti-logaritmo, tem-se que a razão das chances estimadas, definida como razão das chances (*odds ratio*), é expressada em:

$$\widehat{OR} = \frac{chance_2}{chance_1} = \hat{\beta}_1. \tag{2.14}$$

2.4 INFERÊNCIA PARA O MODELO LOGÍSTICO

Após realizar as estimativas dos coeficientes, procura-se avaliar a significância das variáveis do modelo. Será feita a comparação dos valores observados da variável resposta para dois modelos (com e sem a variável independente) com o objetivo de verificar se o modelo que inclui uma determinada variável diz mais sobre a variável resposta do que o modelo sem esta variável. O primeiro método utilizado será o da diferença da soma dos quadrados. Sendo a variação não explicada, soma do quadrado dos resíduos:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

e a variação explicada denotada por:

$$SSR = \left[\sum_{i=1}^{n} (y_i - \overline{y_i})^2\right] - \left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right].$$

Na regressão logística comparam-se os valores observados da variável resposta com os valores preditores dos modelos, com e sem a variável em questão, através do log da função de verossimilhança definida em (2.7). A comparação entre esses valores utilizando a função de verossimilhança se dará por:

$$D = -2ln\left[\frac{(verossimilhaca\ do\ modelo\ ajustado)}{(verossimilhanca\ do\ modelo\ saturado)}\right]. \tag{2.15}$$

Na situação em que os valores da variável resposta são 0 ou 1, a verossimilhança do modelo saturado é 1, onde temos pela definição do modelo saturado que $\hat{\pi}(x_i) = y_i$. Sendo assim:

$$D = -2ln(verossimilhanca\ do\ modelo\ ajustado). \tag{2.16}$$

Então, para avaliar a significância de uma variável, será considerado o valor de D com e sem a tal variável

$$G = D(modelo\ sem\ a\ variavel) - D(modelo\ com\ a\ variavel)$$

que pode ser expressada por:

$$G = -2ln\left[\frac{(verossimilhanca\ do\ modelo\ sem\ a\ variavel)}{(verossimilhanca\ do\ modelo\ com\ a\ variavel)}\right]. \tag{2.17}$$

Partindo da situação em que se tem apenas uma variável independente, a estatística G obedecerá à uma distribuição Qui-Quadrado com 1 grau de liberdade (considerando uma amostra grande de dados). Com isso quantifica-se a significância da variável calculando o $p - valor$ associado à $P[\mathcal{X}^2 > G]$, em que valores pequenos indicam boa significância.

Um segundo teste sugerido em Hosmer and Lemeshow (2000) para verificar a significância da variável independente é o teste Wald, obtido pela comparação entre a estimativa de máxima verossimilhança do parâmetro ($\hat{\beta}_1$) e a estimativa de seu erro padrão. A razão resultante, sob a hipótese H_o: $\beta_1 = 0$, tem distribuição normal padrão. A estatística do teste Wald para a regressão logística é:

$$W_j = \frac{\hat{\beta}_1}{\widehat{EP}(\hat{\beta}_1)}$$

sendo que o $p - valor$ é definido como $P(|Z| > |W_j|)$, Z a variável aleatória da distribuição normal padrão e $\widehat{EP}(\hat{\beta}_1)$ o erro padrão da estimativa de β_1. No entanto, recomenda-se a utilização do teste da razão de verossimilhança para testar se realmente o coeficiente não é significativo quando o teste de Wald não rejeitar a hipótese nula, pois o teste Wald pode se comportar de maneira inadequada em algumas situações.

2.5 REGRESSÃO LOGÍSTICA MÚLTIPLA

Anteriormente foi apresentado o modelo de regressão logística considerando apenas uma variável explicativa. Bem como no modelo de regressão linear, o modelo trabalhado também poderá ser ajustado levando em conta mais de uma variável explicativa, o que o define como um Modelo de Regressão Logística Múltipla. Considerando que o modelo possui um conjunto de p variáveis independentes denotadas por um vetor $X = (X_1, X_2, X_3, ..., X_p)$, então o logito do modelo de regressão múltipla será:

$$\mathbf{g(X)} = ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X_1 + \boldsymbol{\beta_2} X_2 + \ldots + \boldsymbol{\beta_p} X_p. \tag{2.18}$$

Sendo o modelo de regressão logística:

$$E(Y) = \pi(X) = \frac{e^{\mathbf{g(X)}}}{1 + e^{\mathbf{g(X)}}}. \tag{2.19}$$

O método de estimação dos parâmetros usado no caso multivariado será o mesmo da situação univariada, o estimador de máxima verossimilhança. A diferença é que agora $\pi(X)$ é definido como na Equação (2.19). As equações de verossimilhança

podem ser expressas por:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

e

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0$$

onde $j = 1, 2, ..., p$.

A solução das equações acima fornecem estimativas dos parâmetros do modelo utilizando processos interativos análogos ao caso univariado. Obtida essas estimativas, calculam-se então as probabilidades ajustadas do modelo:

$$\hat{\pi_i} = \frac{e^{\hat{\mathbf{g}}(\mathbf{X_i})}}{1 + e^{\hat{\mathbf{g}}(\mathbf{X_i})}} \tag{2.20}$$

sendo $\hat{\mathbf{g}}(\mathbf{X_i})$ definido em (2.18).

Tratando-se agora da estimativa do erro padrão, o método para estimar as variâncias e covariâncias dos coeficientes estimados segue a teoria da estimação de máxima verossimilhança, que assegura que os estimadores são obtidos da matriz de derivadas segundas parciais da função log de verossimilhança, tendo a seguinte forma geral:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1 - \pi_i) \tag{2.21}$$

e

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i(1 - \pi_i) \tag{2.22}$$

onde $j, l = 0, 1, 2, ..., p$ e π_i simplifica $\pi(x_i)$.

Seja a matriz $(p + 1) \times (p + 1)$ que contém os termos negativos de (2.21) e (2.22) denotada por $\boldsymbol{I(\beta)}$: matriz informação de Fisher, tem-se que a partir do

inverso dessa matriz pode-se obter as variâncias e as covariâncias dos coeficientes estimados, definida como $Var(\boldsymbol{\beta}) = \boldsymbol{I^{-1}}(\boldsymbol{\beta})$. A notação $Var(\beta_j)$ retorna o j^{th} elemento da diagonal da matriz e $Cov(\beta_j, \beta_l)$ denota um elemento fora da diagonal que é a covariância de $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ e $\hat{\boldsymbol{\beta}}_{\boldsymbol{l}}$ .Os estimadores das variâncias e covariâncias, $\widehat{Var}(\hat{\boldsymbol{\beta}})$, são obtidos de $Var(\boldsymbol{\beta})$ em $\hat{\boldsymbol{\beta}}$. Os valores da matriz serão definidos por $\widehat{Var}(\hat{\beta}_j)$ e $\widehat{Cov}(\hat{\beta}_j, \hat{\beta}_l)$. A matriz de informação de Fisher estimada pode ser obtida por: $\hat{I}(\hat{\boldsymbol{\beta}}) = \mathbf{X'VX}$.

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix} \tag{2.23}$$

e

$$\mathbf{V} = \begin{bmatrix} m_1\hat{\pi_1}(1-\hat{\pi_1}) & 0 & \dots & 0 \\ 0 & m_2\hat{\pi_2}1(1-\hat{\pi_2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_n\hat{\pi_n}(1-\hat{\pi_n}) \end{bmatrix}. \tag{2.24}$$

No que diz respeito à inferência, bem como o caso de regressão logística simples, a significância dos parâmetros será testada a partir do teste de Razão de Verossimilhança. O teste para a significância dos p coeficientes das variáveis independentes do modelo, é realizado da mesma maneira que em (2.15). No caso da regressão múltipla, tem-se o interesse em saber se pelo menos uma variável é significativa para o modelo. Sob a hipóstese nula, os p coeficientes são iguais a zero, assim, a estatística G tem distribuição Qui-Quadrado com p graus de liberdade.

Capítulo 3

AMOSTRAGEM PARA EVENTOS RAROS

3.1 INTRODUÇÃO

A amostragem possibilita o estudo de um pequeno grupo de elementos retirados de uma população que se pretende conhecer. Trata-se de uma técnica de pesquisa na qual um conjunto pré-estabelecido de amostras é considerado adequado para estimar características de toda a população estudada, com margem de erro definida. No entanto, cada banco de dados possui uma realidade diferente, e com isso deve-se escolher a técnica de amostragem que melhor possibilita a obtenção de resultados fidedignos à população.

O modelo de regressão logística da variável independente binária 0 e 1, ou seja, cliente pagou ou não pagou o crédito concedido, necessita que a proporção dessas classes seja balanceada, do contrário o modelo não será um bom preditor da categoria minoritária. A literatura (Alves and Silva, 2013) considera que se um evento possuir menos de 15% de frequência, então é classificado como evento raro. Nesse caso, um tratamento diferenciado deve ser tomado para a análise dos dados.

As estimativas $\hat{\boldsymbol{\beta}}$ de (2.7) possuem matriz de covariância:

$$V(\hat{\boldsymbol{\beta}}) = \sum_{i-1}^{n} [\pi_i(1-\pi_i)X_i'X_i]^{-1} \tag{3.1}$$

King and Zeng (2001) apontam que, se o modelo logit possui razoável poder de explicação, as probabilidades estimadas serão relativamente próximas de 0,5 para $Y_i = 1$ e mais próximas de zero para $Y_i = 0$. A quantidade $\pi_i(1-\pi_i)$ será maior entre os eventos raros, consequentemente a quantidade $[\pi_i(1-\pi_i)X_i'X_i]^{-1}$ será menor quando $Y_i = 1$. Tal característica indica que a inclusão de mais sucessos na amostra é mais informativa que a inclusão de mais fracassos, sendo assim a técnica de amostragem mais adequada a este tipo de banco de dados será aquela que proporcionar um aumento no número de eventos raros.

Para facilitar a ilustração do problema, a classe minoritária será classificada como positivo e a classe majoritária será classificada como negativo. A Tabela 3.1 ilustra como o modelo preditor pode ser comportar, onde verdadeiro positivo (VP) e verdadeiro negativo (VN) denotam o número de eventos $y_i = 1$ e $y_i = 0$,respectivamente, que são classificados corretamente enquanto FP e FN significam erro na classificação dos eventos positivo e negativo, respectivamente.

Tabela 3.1: Tabela de classificação dos eventos

	Preditor Positivo ($\hat{y}_i = 1$)	Preditor Negativo ($\hat{y}_i = 0$)
Real Positivo ($y_i = 1$)	VP	FN
Real Negativo ($y_i = 0$)	FP	VN

A acurácia mede o quanto a estimativa que obtivemos é relacionada com o "valor real" do parâmetro. Ela nos informa o quanto o valor estimado é "bom", ou seja,

quanto o valor estimado é próximo do valor real. Ela é calculada da seguinte forma:

$$Acuracia = (VP + VN)/(VP + FN + FP + VN) \tag{3.2}$$

Quando usa-se a acurácia para medir a performance do modelo, ela estará apta a predizer a classe dominante melhor que a classe rara. Essa conclusão pode ser verificada ao se analisar a Equação 3.2 onde, se a base de dados é extremamente desbalanceada, mesmo quando o classificador classificar toda a classe rara de forma errada, a acurácia continuará alta se a classe dominante tiver predição correta porque existe muito mais eventos $y_i = 0$, sobre essas circunstâncias a acurácia não consegue refletir uma predição confiável para a classe rara.

Este trabalho propõe utilizar as técnicas de reamostragem e de imputação de observações sintéticas SMOTE para corrigir a problemática causada por estes eventos raros.

3.2 EVENTOS RAROS

É comum, nas mais diversas áreas de conhecimento, a variável resposta de interesse possuir distribuição dicotômica extremamente desbalanceada. No mercado financeiro a problemática de base de dados com eventos raros é evidenciada pela frequência extremamente pequena de clientes fraudulentos.

Existem alguns estudos que revelam que o modelo de regressão logística usual subestima a probabilidade dos eventos de interesse quando este é construído utilizando base de dados extremamente desbalanceada devido ocorrência de eventos raros (King and Zeng, 2001). Em Greene (2008) aponta-se que as funções de ligação

logística produzem resultados distintos em amostras com baixa frequência de "sucesso" ($Y_i = 1$ *cliente inadimplente*) em relação ao número de "fracasso" ($Y_i = 0$ *cliente adimplente*).

Como dito anteriormente, a média da variável binária é a frequência relativa dos eventos, e por isso uma base de dados desbalanceada traz consequências importantes para as análises produzidas.

3.3 TÉCNICA SMOTE

Algoritmos classificadores são sensíveis ao desbalanceamento, e tendem a supervalorizar o evento predominante e muitas vezes a ignorar os eventos de menor frequência. Segundo Machado and Ladeira (2007) a técnica de *oversampling* -sobreamostragem- puramente não é bem aceita na comunidade científica, pois em muitos dos casos estas técnicas apenas reproduzem casos existentes. Neste estudo também é considerado que esse tipo de replicação aumenta o víeis do classificador e que, além disso, acontece um efeito indesejado de modelos *overfitted* -super ajustados-, em que os modelos ficam muito específicos para os casos replicados, prejudicando seu poder de generalização para a classe de interesse. Replicar meramente os casos de menor frequência possibilita que os classificadores reconheçam a região, no entanto tal região será tão pequena que não conseguirá classificar corretamente novos casos da classe de interesse que venham a cair nas vizinhanças desta região.

Para reduzir o viés presente nas estimativas de bases de dados com classes minoritárias, não deixando os classificadores serem afetados pela problemática do *oversampling* mencionada, Chawla et al. (2002) sugeriram um método computacional

que consiste na geração de casos sintéticos (imputação de observações artificiais) para a classe de interesse a partir dos casos já existentes. Estas observações são geradas na vizinhança de cada caso de evento raro, de forma a fazer crescer a região de decisão. Esta nova técnica denominada pelos autores de SMOTE é um algoritmo que possibilita criar amostras sintéticas a partir da classe com poucas observações por meio de um pseudocódigo para geração de amostra sintética. O método consiste basicamente em, a partir de cada observação original de evento raro, gerar aleatoriamente uma observação sintética ao longo do segmento de reta que une a observação de evento raro com seus k vizinhos mais próximos aleatoriamente escolhidos. O número k de vizinhos será definido de acordo com o quanto se deseja aumentar a classe rara trabalhada. Caso seja necessário aumentar em 300%, por exemplo, então será necessário aplicar o algoritmo para $k = 3$ vizinhos, gerando uma observação sintética na direção de cada um desses vizinhos.

Por Rocha and Eirado (2012), a observação sintética pode ser calculada da seguinte forma:

$$obs_n = obs_i + ale * dif \tag{3.3}$$

onde:

obs_n = nova observação sintética;

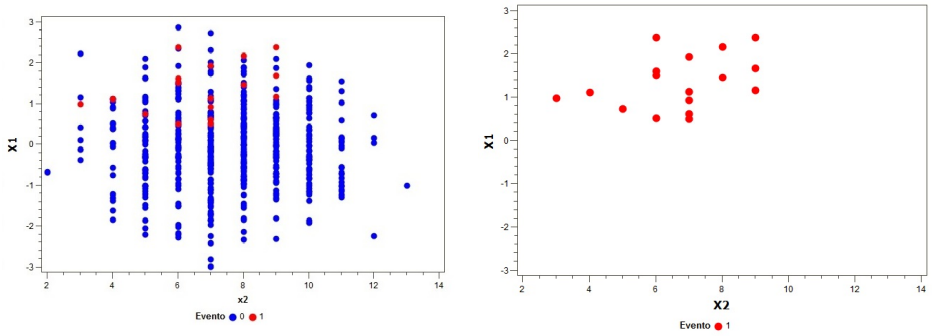
obs_i = i-ésima observação do evento raro, selecionada aleatoriamente sem reposição ;

obs_j = j-ésima observação do evento raro, selecionada aleatoriamente entre os k vizinhso mais próximos de obs_i;

$dif = obs_j - obs_i$ = diferença entre a i-ésima e a j-ésima observação;

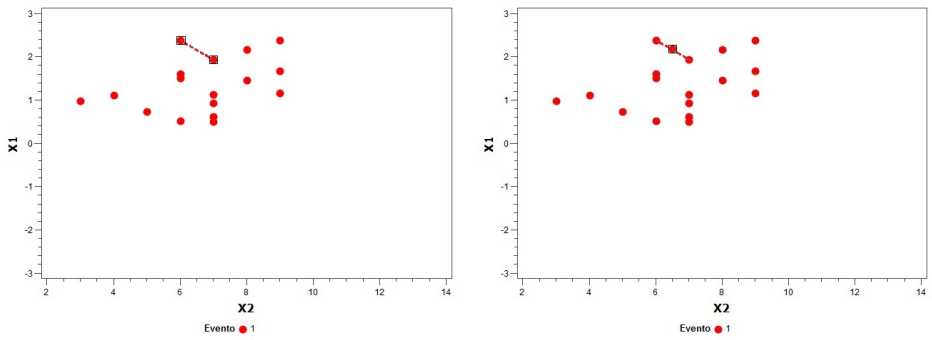
ale = um número aleatório entre (0,1).

Em que se calcula a diferença do vetor característico da observação obs_i e do seu vizinho obs_j, multiplica-se essa diferença por um valor aleatório entre 0 e 1, e se adiciona esse valor ao vetor característico de obs_i. Desta forma cria-se uma nova observação que será um ponto aleatório no segmento de reta que liga essas duas observações. A imputação dessas observações na amostra original fará com que a região de decisão do evento raro se torne mais geral, possibilitando um maior percentual de acerto na predição. As Figuras 3.1 e 3.2 exemplificam a criação da observação sintética em uma base simulada de 500 observações.



(a) Simulação 1 (b) Seleção de $y_i = 1$

Figura 3.1: Seleção de evento raro



(a) Seleção de observação e seu vizinho (b) Imputação de observação sintética

Figura 3.2: SMOTE

3.4 TÉCNICA DE REAMOSTRAGEM

O procedimento proposto em Alves and Silva (2013) baseia-se no ajuste de um modelo probit feito a partir da seleção de todas as m unidades amostrais pertencentes ao evento raro ($Y_i = 1$) e seleção aleatória sem reposição de m unidades do evento de frequência predominante ($Y_i = 0$).

No presente estudo propõe-se utilizar a mesma metodologia sugerida por Alves and Silva (2013), agora para o caso de ajuste do modelo *logit*. O método de reamostragem utilizado resultará em sub-amostras compostas de 50% observações pertencentes ao evento raro ($Y_i = 1$, clientes inadimplentes) e de 50% observações pertencentes ao evento predominante ($Y_i = 0$, clientes adimplentes). O ajuste do modelo *logit* será feito armazenando, para cada uma das observações, as probabilidades preditas de sucesso $\hat{p}_{ib}(X_b)$. Repete-se este procedimento B vezes até que todas as unidades observacionais do grupo de maior frequência sejam selecionadas ao menos uma vez. Após este processo, as médias das probabilidades preditas são calculadas para cada uma das observações.

$$b = 1, ..., B \Rightarrow \begin{cases} P(X_{bi} = 1) = \pi(X_b'\beta_b) \\ P(X_{bi} = 0) = 1 - \pi(X_b'\beta_b) \end{cases} \tag{3.4}$$

$$\hat{p}_{ib}(X_b) = [Y_{bi} \times \pi(X_b'\hat{\beta}_b)] + [(1 - Y_{bi}) \times (1 - \pi(X_b'\hat{\beta}_b))] \tag{3.5}$$

$$\overline{\hat{p}_i}(X) = \sum_{b-1}^{B} \frac{\hat{p}_{ib}(X_b)}{B} \tag{3.6}$$

Ajustando esses B modelos logit em sub-amostras de 50% inadimplentes e 50% adimplentes, contorna-se o problema de excesso de zeros na amostra. A aleatoriedade na atribuição dos zeros, selecionadas repetidas vezes através de amostragem sem

reposição, contorna a problemática de seleção viesada, a função de verossimilhança não será mais demasiadamente influenciada pela grande quantidade de fracassos. Em cada uma das sub-amostras o modelo probabilístico consegue discernir sucessos dos fracassos no momento de construir as probabilidades preditas.

Este método pode ser considerado relativamente simples, podendo facilmente ser realizado com auxílio do procedimento SURVEYSELECT do *software* SAS. Entretanto, um dos impasses presentes nessa metodologia é identificado quando se trabalha com amostras muito grandes, problema este recorrente nas bases bancárias, tornando o processo computacionalmente intensivo, pois é necessário que todas as unidades observacionais possuam ao menos uma probabilidade de sucesso estimada. Sendo assim, será necessário ajustar um número muito grande de modelos probabilísticos, de tal forma que toda população seja preenchida.

Capítulo 4

MATERIAL E MÉTODOS

4.1 INTRODUÇÃO

Neste capítulo serão detalhados os procedimentos realizados durante o estudo, com descrição de técnicas e metodologia utilizadas na criação da base de dados simuladas e na aplicação das técnicas de reamostragem e $SMOTE$.

4.2 MATERIAL

Com o auxílio do Software SAS gerou-se 50 simulações de 500 observações e 50 simulações de 10.000 observações para diferentes intensidades do processo de auto-seleção (a), onde $0 < a < 1$ como proposto em Alves and Silva (2013). A auto-seleção de indivíduos ocorre devido aos pré-requisitos usualmente solicitados pelos bancos, tais como renda, idade, se possui imóvel, etc. Neste estudo serão simulados diferentes graus de auto-seleção, que representam as diferentes porcentagens de eventos raros.

Considerando a variável y_i de distribuição Bernoulli com probabilidade de sucesso $\pi_i(y_i = 1)$ e a variável U_i^* com distribuição uniforme ($U_i^* \sim U(a, 0)$), então y_i determina cada uma das ocorrências binárias segundo o esquema:

$$\begin{cases} \pi(y_i) \leq U_i^* \Rightarrow y_i = 0 \\ \pi(y_i) > U_i^* \Rightarrow y_i = 1 \end{cases} \tag{4.1}$$

Se $a = 0$, então U_i^* tem distribuição $U_i^* \sim U(0,1)$ e todas as unidades observacionais possuem alguma chance de sucesso, inclusive aquelas que possuem probabilidade nula ($\pi(y_i) = 0$). As diferentes intensidades do processo de auto-seleção são efetuadas variando-se $0,00 < a < 0,99$. Por exemplo, para $a = 0,99$ temos um processo de auto-seleção, onde somente 1% das observações possui alguma chance de sucesso ($\pi(y_i) \leq 0,99$). Seguindo a sugestão de Alves and Silva (2013), será utilizado uma equação contendo duas variáveis explicativas (x_{i1} e x_{i2}), entretanto o esquema de simulação pode ser facilmente estendido para um modelo com mais variáveis. As variáveis explicativas possuem distribuição de probabilidade, respectivamente, normal e binomial: $x_{i1} \sim N(0;1)$ e $x_{i2} \sim BIN(15;0,5)$. Foram escolhidos os seguintes tamanhos de amostra:

$$n = \{500, 10.000\} \tag{4.2}$$

O intercepto possui valor fixo $\beta_0 = 8,75$ e os coeficientes angulares associados a x_{i1} e x_{i2} possuem valores também fixos, respectivamente, 4 e $-1,17$. Estes valores foram baseados no modelo definido por (Alves and Silva, 2013)

$$Y_i = 8,75 + 4x_{i1} - 1,17x_{i2} \tag{4.3}$$

Para a Equação (4.3) e para cada um dos tamanhos de amostras (4.2) serão simuladas 20 amostras contendo os seguintes processos de auto-seleção:

$$a = \{\ 0;\ 0,05; 0,10;\ \ldots\ ;\ 0,90;\ 0,95\ \} \tag{4.4}$$

Retomando a motivação de análise de concessão de crédito, para $a = 0$ todas os indivíduos possuem chance de terem seu crédito concedido pelo banco. Para

$a = 0,99$ somente os 1% dos considerados melhores clientes serão analisados para concessão de crédito. A simulação do processo de auto-seleção encontra-se representado pela expressão:

$$\begin{cases} \pi(8,75 + 4x_{i1} - 1,17x_{i2}) \leq U_i^* \sim U(a,1) \Rightarrow y_i = 0 \\ \pi(8,75 + 4x_{i1} - 1,17x_{i2}) > U_i^* \sim U(a,1) \Rightarrow y_i = 1 \end{cases} \tag{4.5}$$

Este procedimento torna possível avaliar o efeito da intensidade do processo de auto-seleção dado diferentes tamanhos de amostra. A probabilidade real de sucesso é conhecida sendo possível também se obter o Erro Quadrático Médio (EQM) definido como a forma de avaliar a variância e o viés do estimador, sendo que o EQM mínimo indicará a variação mínima e portanto indicará o melhor estimador.

$$EQM(p, a; n) = E(p_{an} - \hat{p}_{an})^2 \tag{4.6}$$

4.3 MÉTODOS

Serão feitas simulações de diferentes intensidades de evento raro nos diferentes tamanhos de amostra, em seguida aplicada a técnica de reamostragem definida na Seção 3.4 e a técnica de imputação de observações definida na Seção 3.3, esta última inicialmente para $k = 1$ onde espera-se aumentar em 100% a frequência de eventos raros. Caso o percentual de acerto não melhore, então será ajustado um k de acordo com a necessidade de cada percentual de auto-seleção. A avaliação será feita através da predição das ocorrências binárias entre as observações, baseando-se para isto nas probabilidades preditas estimadas. Uma verificação simples pode ser realizada adotando-se a regra :

$$\begin{cases} \pi(\hat{\beta_{0b}} + \hat{\beta_{1b}}x_{i1} - \hat{\beta_{2b}}x_{i2}) < 0,5 \Rightarrow \hat{y}_i = 0 \\ \pi(\hat{\beta_{0b}} + \hat{\beta_{1b}}x_{i1} - \hat{\beta_{2b}}x_{i2}) \geq 0,5 \Rightarrow \hat{y}_i = 1 \end{cases} \tag{4.7}$$

Caso tenham sido produzidas boas estimativas para as probabilidades, se espera uma alta concentração relativa nos pontos ($y_i = 0; \quad \hat{y}_i = 0$) e ($y_i = 1; \quad \hat{y}_i = 1$). Espera-se que em processos de auto-seleção caracterizados por baixos valores de a, o ajuste de um único modelo probabilístico produza melhores resultados do que a simulação proposta. Para processos de auto-seleção caracterizados por altos valores de a, espera-se que as técnicas de amostragem propostas apresentem melhores resultados (Alves and Silva, 2013).

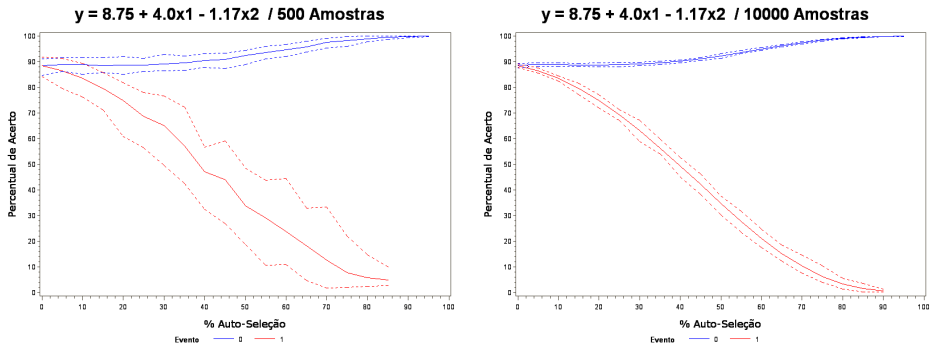
Capítulo 5

ANÁLISE DE RESULTADOS

5.1 INTRODUÇÃO

Nesta seção são apresentados os resultados gerados a partir das bases simuladas como definido na Seção 4.2.

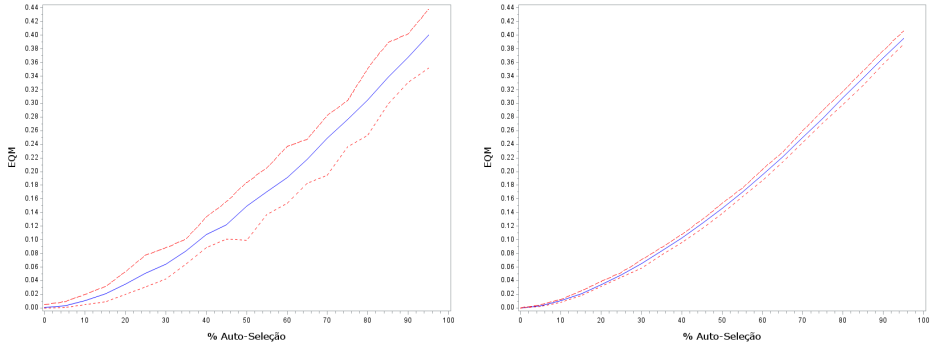
A Figura 5.1 evidencia o problema de desbalanceamento da base de dados, em que há baixo percentual de acerto para altos graus de auto-seleção, tanto para o caso de simulaçao de 500 quanto para de 10.000 observações. Para ilustração, também apresenta-se os mínimos e máximos das estimativas representados pelas linhas traçejadas. Reforça-se, com as figuras, a necessidade de se utilizar técnicas que possibilitem que o modelo continue um bom preditor, por mais que o evento diminua sua frequência.



(a) 500 Observações (b) 10.000 Observações

Figura 5.1: Percentual de Acerto x Auto-Seleção

Já na Figura 5.2 apresenta-se a evolução do Erro Quadrático Médio nos diferentes valores de auto-seleção, é visível que os EQMs apresentam comportamento estritamente crescente, pois a medida que se aumenta a auto-seleção maior será o erro do estimador. As técnicas de amostragem propostas nesse trabalho visam diminuir o viés de estimadores dessas classes denominadas raras.

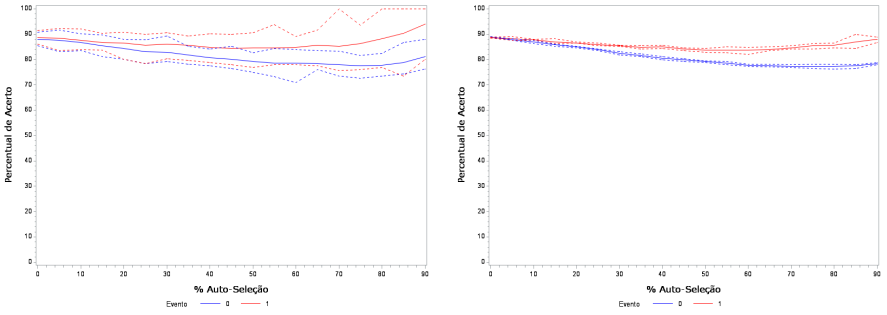


(a) 500 Observações (b) 10.000 Observações

Figura 5.2: EQM x Auto-Seleção

5.2 REAMOSTRAGEM

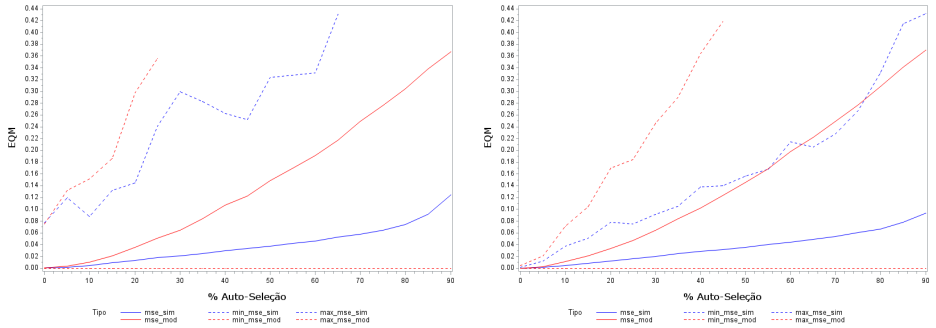
Após aplicada a técnica de reamostragem nas bases simuladas, espera-se que ela equilibre o percentual de acerto do modelo para os diferentes graus de auto-seleção, reduzindo assim as consequências negativas de uma base de dados desbalanceada. A Figura 5.3 evidencia esses resultados equilibrados.



(a) 500 Observações (b) 10.000 Observações

Figura 5.3: Percentual de Acerto x Auto-Seleção

Na Figura 5.4 são apresentados os resultados da comparação do modelo logit com a técnica de reamostragem, sendo que a linha vermelha contínua mostra a comparação do EQM da regressão logit ajustado para toda a população, e a linha azul contínua mostra o EQM segundo a metodologia apresentada na Seção 3.4. O eixo horizontal mostra diferentes intensidades do processo de auto-seleção, enquanto o eixo vertical mostra o EQM, e as respectivas linhas pontilhadas representam os limites inferiores e superiores.



(a) 500 Observações (b) 10.000 Observações

Figura 5.4: EQM x Auto-Seleção

Considerando a amostra de tamanho 500, o método de simulação proposto se comporta melhor que o modelo logit para todas as intensidades do processo de auto-seleção ($0 < a < 1$), evidenciando EQM abaixo para todos os valores do eixo. Onde, quanto maior o grau de auto-seleção, ou seja, menor a frequencia de $Y_i = 1$ (evento raro), melhor o método de reamostragem se comporta comparado ao modelo.

5.3 SMOTE

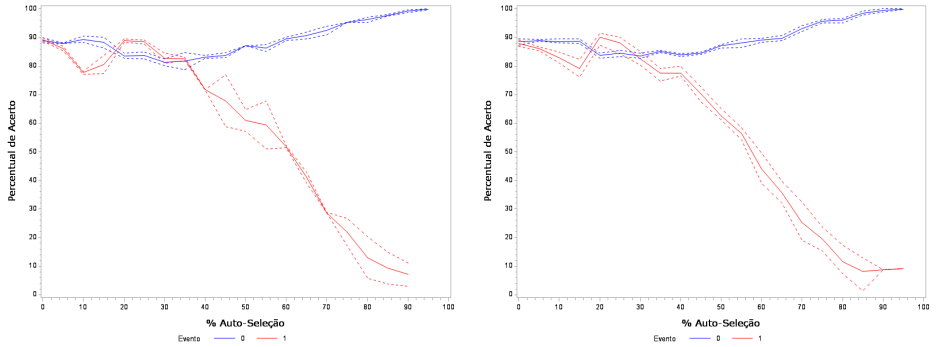
5.3.1 SMOTE para $k = 1$

A técnica SMOTE, aplicada para aumentar em 100% a classe de menor

frequência, não trouxe resultados equilibrados para os percentuais de auto-seleção maiores, seu uso não se mostrou eficaz para a problemática de eventos raros, pois aumentar em 100% ($k = 1$) uma base em que a classe de menor frequência tem representatividade de apenas 5%, significa aumentar para aproximadamente 10% sua frequência. Dessa forma sua representatividade continua baixa e os resultados gerados continuarão com baixo percentual de acerto para $y_i = 1$.

Exemplificando para fins de fácil entendimento: caso uma base de 200 observações fosse composta por 190 indivíduos $y_i = 0$ e 10 indivíduos $y_i = 1$, com a aplicação da técnica SMOTE para $k = 1$, essa base mudaria para $190 : y_i = 0$ e $20 : y_i = 1$, como pode-se observar a classe minoritária continua muito menor que a predominante, sendo assim o desbalanceio não foi corrigido.

Esses resultados mencionados podem ser verificados na Figura 5.5, onde a técnica SMOTE conseguiu manter um bom percentual de acerto apenas até o grau de auto-seleção de aproximadaemtne 35%.

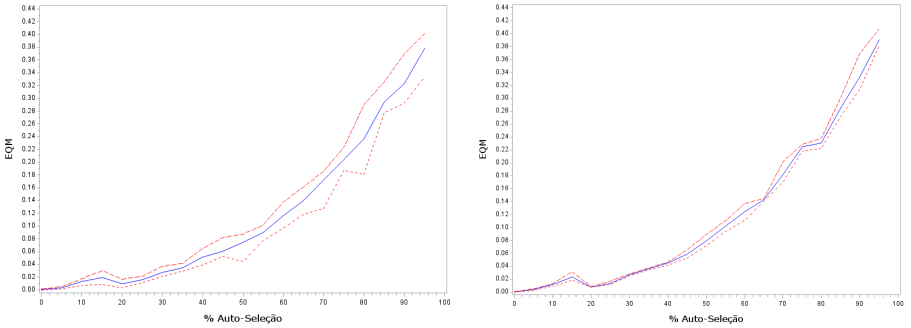


(a) 500 Observações (b) 10.000 Observações

Figura 5.5: Percentual de Acerto x Auto-Seleção

A Figura 5.6 evidencia o mesmo problema, o EQM apresentou resultado estritamente crescente, a inclinação foi mais suave se comparada a base que não foi tratada

com nenhuma técnica, no entanto os resultados não foram expressivamente melhores já que os dados continuaram desbalanceados.



(a) 500 Observações (b) 10.000 Observações

Figura 5.6: EQM x Auto-Seleção

5.3.2 SMOTE AJUSTADO

A técnica SMOTE aplicada para $k = 1$, em que se aumenta em 100% a classe de menor frequência, não foi suficiente para balancear os dados, para fim de igualar as frequência das duas classes buscou-se aumentar o número de $y_i = 1$ até que ele atingisse o número de $y_i = 0$, dessa forma aumentou-se o número de vizinhos k selecionados por observação original, de acordo com a necessidade de balanceio de cada grau de auto-seleção. Esta técnica ajustada para cada viés foi aplicada apenas para as simulações de 500 observaçoes, dado que a utilização deste ajuste nas simulações de 10.000 observações se mostrou excessivamente extensiva. A Figura 5.7 mostra um bom percentual de acerto utilizando esta metodologia.

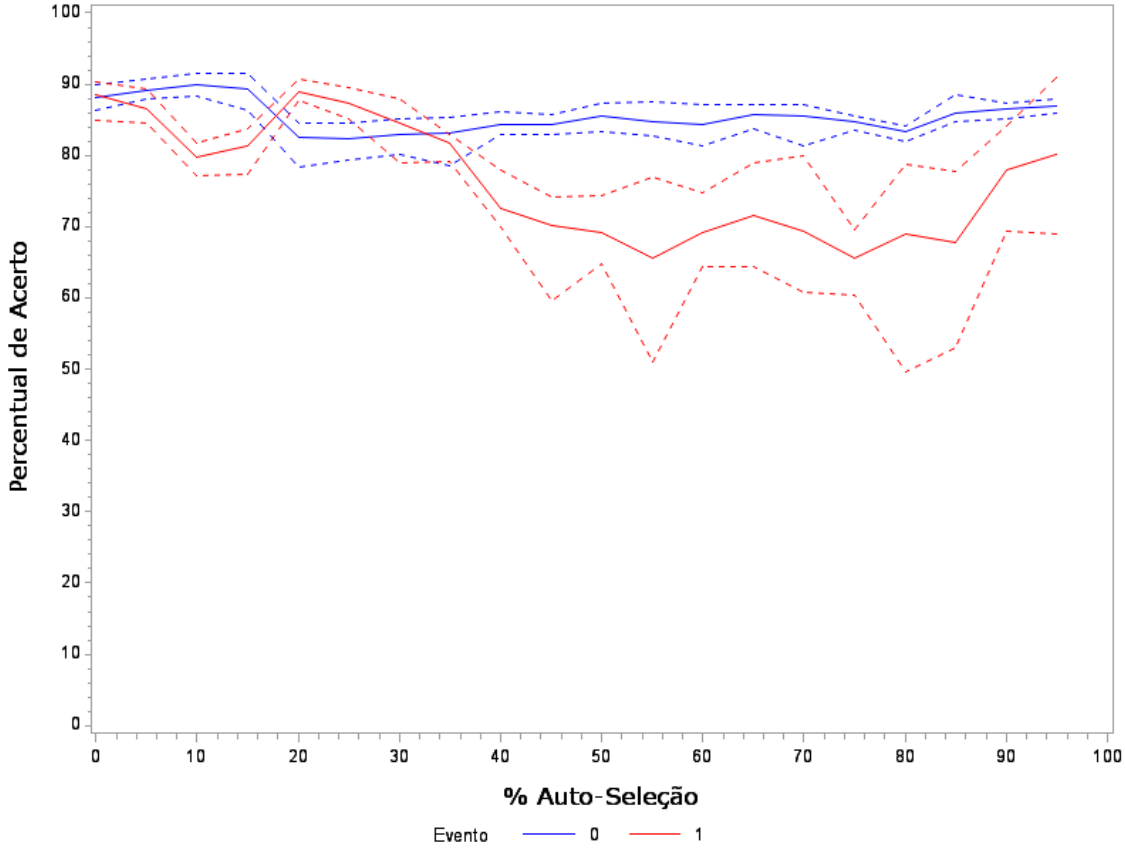


Figura 5.7: Percetual de Acerto x Auto-Seleção para 500 Observações

5.3.3 COMPARAÇÃO DAS TÉCNICAS

A Figura 5.8 compara o percentual de acerto da técnica de Reamostragem e da técnica SMOTE Ajustado com a base original, de tal forma que se pode verificar como cada uma se comporta nos diferentes percentuais de desbalanceio.

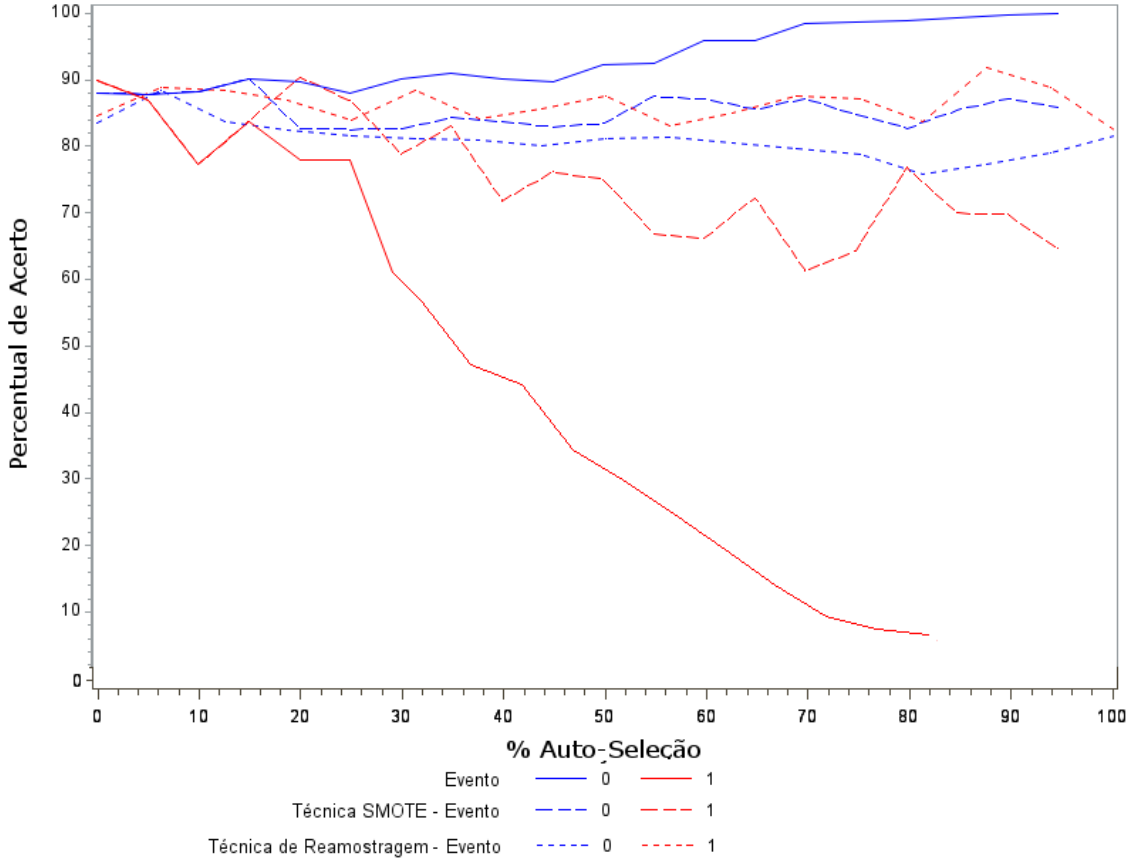


Figura 5.8: Percetual de Acerto x Auto-Seleção para 500 Observações

Para bases mais desbalanceadas, que foi a motivação para a realização deste trabalho, a técnica de Reamostragem e a de SMOTE conseguiram melhorar consideravelmente o percentual de acerto do modelo. No entanto, a Reamostragem fez essa melhora sem precisar criar nenhuma observação sintética, ao contrário do que foi feito no SMOTE, com isso seus dados permaneceram mais fidedignos à base original e, consequentemente, as estimativas da Reamostragem irão possuir menos viés.

Capítulo 6

CONCLUSÃO

Analisando os resultados obtidos, a técnica de Reamostragem se mostrou adequada para tratar bases de dados desbalanceadas. Os subgrupos, formados para balancear os dados, possibilitaram o equilíbrio no percentual de acerto e uma diminuição significativa do Erro Quadrático Médio principalmente na ocorrência de eventos raros, evidenciando uma melhora consistente no modelo após a utilização da metodologia proposta na Seção 3.4.

No que diz respeito ao procedimento SMOTE, quando ajustado para cada grau de auto-seleção, sua aplicação tornou o modelo um bom preditor da classe minoritária, equiparando o percentual de acerto de $y_i = 0$ e $y_i = 1$, e com isso tornou-se possível que o modelo gerasse análises fidedignas a ambas as classes. Entretanto, a utilização de uma técnica que cria observações traz suas ressalvas, pois a base gerada produz estimativas menos fidedignas às suas observações reais.

Por produzir percentuais de acertos tão bons quantos a técnica SMOTE mas sem suas limitações, a Reamostragem se mostrou mais adequada no tratamento de bases de dados desbalanceadas. Após utilizada a metodologia, houve uma melhora no desempenho do modelo de tal forma que tornou viável a análise de eventos raros,

o modelo que antes possuía uma boa acurácia apenas para a classe de frequência predominande, agora tournou-se um bom preditor também para a classe de baixa frequência.

Referências Bibliográficas

Alves, P. F. & Silva, A. R. (2013). Modelagem de eventos raros: Uma aplicação utilizando regressão probit. *Submetido para publicação*.

Chambers, E. A. & Cox, D. R. (1967). Discrinimation betwenn alternative binary response models. *Biometrika*, 54:573–578.

Chawla, N. V., Bowyer, K. W., & Hall, L. O. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:p. 321 – 357.

Fernandes, G. & Rocha, C. A. (2011). Low default modelling: a comparison of techniques based on a real brazilian corporate portfolio. Technical report.

Greene, W. W. (2008). *Econometric Analysis*. Prentice Hall.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley-Interscience Publication.

King, G. & Zeng, L. (2001). Logistic regression in rare event data. *Political Analysis*, 9(2):137–163.

Machado, E. L. & Ladeira, M. (2007). Um estudo de limpeza em base de dados desbalanceada com sobreposição de classes. Technical report.

Rocha, L. C. S. & Eirado, C. R. (2012). Smote: Synthetic minority over-sampling technique for low-default portfolios. Technical report.

SAS (2011). *SAS on line doc, Versão 9.3*. Cary, NC: SAS Institute Inc.