UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM MODELOS DE *CREDIT SCORING*

MARIANA NOLDE PACHECO

Porto Alegre
2016

MARIANA NOLDE PACHECO

MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM MODELOS DE *CREDIT SCORING*

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Orientadora
Professora Dra. Lisiane Priscila Roldão Selau

Porto Alegre
2016

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM MODELOS DE *CREDIT SCORING*

MARIANA NOLDE PACHECO

Banca examinadora:

Professora Dra. Lisiane Priscila Roldão Selau
UFRGS

Bel. Marcos Roberto Eilert de Souza
BANCO COOPERATIVO SICREDI

AGRADECIMENTOS

À UFRGS, pela oportunidade de ensino, aprendizado, trabalho e crescimento pessoal e profissional.

Aos professores do Instituto de Matemática e Estatística da UFRGS por todo o aprendizado e dedicação oferecidos nesses quatro anos de graduação.

Aos amigos que tive oportunidade de conhecer e conviver durante toda a faculdade. Sem o apoio e ajuda de vocês jamais seria possível percorrer toda essa trajetória.

Aos meus colegas de trabalho pela oportunidade, confiança e condições para a realização desse estudo. Um agradecimento em especial para Mariana Mizutani e Marcos Eilert pela sugestão de estudo, pelo incentivo e pelo apoio durante todo esse processo.

Aos meus amigos e familiares por entenderem os momentos de ausência e me incentivarem tanto.

Ao amor incondicional dos meus pais, que entenderam e aceitaram todas as minhas escolhas com dedicação, incentivo, carinho e confiança durante toda a minha vida. Sem o auxílio de vocês nada disso seria possível.

Ao meu noivo e amor da minha vida, Adriano Detoni Filho, responsável pelo meu ingresso no curso de Estatística. Obrigada pelo amor, dedicação, generosidade e paciência ao longo desses anos. Agradeço principalmente por acreditar em mim e me incentivar nas decisões mais importantes da minha vida.

Para finalizar, o agradecimento para minha amiga e orientadora Lisiane Selau, pelo exemplo de dedicação, profissionalismo e amizade. Tuas orientações foram responsáveis por despertar o amor por esse trabalho e área de aplicação da Estatística. Muito obrigada por acreditar e confiar em mim durante todo esse processo.

“Seu trabalho vai preencher boa parte da sua vida e a única maneira de ser verdadeiramente satisfeito é fazer o que acredita ser um ótimo trabalho. E a única maneira de fazer um ótimo trabalho é amar o que você faz.”

Steve Jobs, 2005.

RESUMO

Nos últimos anos, houve aumento na demanda e popularização do mercado de crédito no Brasil. A concessão de crédito envolve riscos, o que pode significar um grande prejuízo monetário para as empresas. Sendo assim, surgiram os modelos de crédito, que buscam identificar características que diferenciam o bom e o mau pagador. Os modelos de *Credit Scoring* são diferenciados conforme a etapa do ciclo de crédito do cliente, sendo divididos geralmente em *Application Scoring, Behavioral Scoring* e *Collection Scoring*. Esses modelos são geralmente construídos com base em uma grande quantidade de características (variáveis) dos clientes, pois podem utilizar informações cadastrais, de crédito, de conta corrente e outras. Dessa forma, é necessário um processo refinado de extração e seleção das variáveis mais importantes na explicação do bom e mau pagador. Existem diversas técnicas de mineração de dados que realizam redução da dimensão de dados originais e/ou seleção de variáveis para utilização em modelos estatísticos que incluem a Análise de Componentes Principais e o método *Stepwise*. Embora amplamente utilizados, nenhum desses dois métodos de seleção de variáveis considera as medidas de desempenho práticas dos modelos (como o resultado do teste KS). Dessa forma, surge a necessidade de verificação da utilização dessas medidas como forma de seleção de variáveis para os modelos de crédito. Sendo assim, o objetivo do presente estudo é comparar modelos de *Credit Scoring* utilizando diferentes métodos de seleção de variáveis: PCA, *Stepwise* e um novo método de seleção baseado no resultado do teste KS, denominado como método Eilert. As informações utilizadas no estudo são provenientes de uma base de dados reais, com duas composições amostrais (desenvolvimento e validação), de um total de 240.000 clientes. Os métodos de seleção foram testados com a utilização de 90 variáveis de cadastro e comportamento dos clientes na empresa. Os modelos finais de crédito foram construídos com a técnica de Regressão Logística, e as medidas de desempenho utilizadas para comparação dos resultados foi o percentual de acerto, o resultado do KS e a curva ROC. Na comparação dos modelos de crédito, construídos com as variáveis indicadas pelos três métodos de seleção, verifica-se que os desempenhos dos modelos, tanto na amostra de desenvolvimento quanto de validação, foram semelhantes, com valores de KS em torno de 57%, ROC em torno de 0,85 e percentual de acerto por volta de 93%. Após a seleção de variáveis com os métodos Eilert, PCA e *Stepwise* com significância de 0,05 e 0,01 os modelos finais apresentaram respectivamente 22, 20, 63 e 56 variáveis. Diante disso, observa-se que os três métodos de seleção de variáveis foram eficazes na redução da dimensão final de variáveis aplicadas nos modelos de *Credit Scoring* construídos no estudo. Sendo assim, a definição do método adotado deve ser considerada através da facilidade de análise, interpretação e apresentação dos resultados dos modelos.

Palavras-chave: Seleção de variáveis. *Credit Scoring*. Regressão Logística.

ABSTRACT

In the last years, there has been an increase in the demand and popularization of the credit market in Brazil. The granting of credit involves risks, which can mean a great deal of monetary loss to companies. Thus, credit models emerged, which seek to identify characteristics that differentiate the good and the bad payer. Credit Scoring models are differentiated according to the stage of the customer's credit cycle, and are usually divided into Application Scoring, Behavioral Scoring and Collection Scoring. These models are usually constructed based on a large number of characteristics (variables) of the clients, since they can use cadastral information, credit, checking account and others. Thus, a refined process of extraction and selection of the most important variables in the explanation of good and bad payer is necessary. There are several data mining techniques that perform the reduction of the original data dimension and / or selection of variables for use in statistical models that include the PCA and the Stepwise method. Although widely used, neither of these two methods of variable selection considers the practical performance measures of the models (such as the KS test result). Thus, there is a need to verify the use of these measures as a way of choosing and selecting variables for credit models. Thus, the objective of the present study is to compare Credit Scoring models using different methods of variable selection: PCA, Stepwise and a new selection method based on the KS test result, denominated Eilert method. The information used in the study comes from a real database, with two sample compositions (development and validation), of a total of 240,000 clients. The selection methods were tested using 90 variables of customer registration and behavior in the company. The final credit models were constructed using the RL technique, and the performance measures used to compare the results were the percentage hit, the KS result and the ROC curve. In the comparison of the credit models constructed with the variables indicated by the three selection methods, it was verified that the performance of the models, both in the development sample and in the validation were similar, with KS values around 57%, ROC around of 0.85 and percentage of accuracy around 93%. However, after selecting variables with the Eilert, PCA and Stepwise methods with significance of 0.05 and 0.01, the final models presented 22, 20, 63 and 56 variables, respectively. Therefore, it is observed that the three methods of variable selection were effective in reducing the final dimension of variables applied in the Credit Scoring models constructed in the study. Therefore, the definition of the adopted method must be considered through the ease of analysis, interpretation and presentation of the results of the models.

Keywords: Selection of variables. Credit Scoring. Logistic Regression

SUMÁRIO

1. INTRODUÇÃO

Diante do crescente desenvolvimento econômico do Brasil nos últimos anos, houve um aumento da demanda e concessão de crédito à população. Segundo Brito (2008), a concessão de crédito consiste na atividade de colocar um valor à disposição de um tomador de recursos, com o compromisso do pagamento do valor emprestado em um determinado período de tempo previamente estabelecido. Dessa forma, conceder crédito envolve riscos, uma vez que é uma decisão tomada em condições de incertezas, onde há a possibilidade de não cumprimento das obrigações financeiras estabelecidas.

De acordo com Steiner *et al*. (1999), a análise correta da concessão de crédito é essencial para a sobrevivência de empresas e instituições financeiras. Erros na concessão de crédito podem significar enormes prejuízos financeiros dentro de uma única operação, gerando a perda do ganho obtido em outras diversas operações de crédito bem-sucedidas. Dessa forma, segundo Sicsú (2010), tem aumentado a busca por diferentes formas de identificação e diferenciação dos bons e maus pagadores. Com isso, há minimização do prejuízo financeiro obtido com transações malsucedidas bem como um acréscimo na rentabilidade (lucro) das instituições.

Diversas formas de concessão de crédito foram sendo aplicadas nas empresas e instituições financeiras ao longo do tempo. Inicialmente, a avaliação dos clientes era realizada por analistas de crédito, profissionais especializados em crédito que observavam as informações dos clientes para decisão de aprovação ou negação das propostas de crédito. Diante da subjetividade dessas avaliações, uma vez que a decisão de crédito variava de acordo com a experiência do analista, surgiram os modelos quantitativos de análise de crédito, também denominados como *Credit Scoring*. Esses modelos são baseados em ferramentas de avaliação quantitativas para classificação do risco de crédito de pessoas físicas (PF) ou jurídicas (PJ) e apresentam como objetivo principal a identificação prévia do bom e mau pagador, reduzindo as transações financeiras equivocadas (MANFIO, 2007; SICSÚ, 2010).

Depois da aprovação e concessão de crédito, os clientes podem se tornar bons ou maus pagadores, de acordo com o comportamento de pagamento. Os bons pagadores geram lucro financeiro, e os maus pagadores são aqueles que geram

inadimplência (não realizam o pagamento do acordo estabelecido durante um determinado período de tempo). Se a inadimplência persistir, os acordos são contabilizados pelas empresas como perda, quando não há mais a previsão de recebimento do valor financeiro emprestado (MANFIO, 2007; SICSÚ, 2010).

Existem diversas técnicas estatísticas que podem ser utilizadas na análise quantitativa de crédito para cálculo dos escores de crédito que fazem a predição de inadimplência dos clientes, tais como: Regressão Linear, Regressão Logística, Análise Discriminante, Redes Neurais Artificiais e outras. Uma das técnicas estatísticas mais conhecidas e utilizadas para avaliação do risco de crédito é a Regressão Logística, que analisa o efeito de uma ou mais variáveis explicativas (categóricas ou métricas) sobre uma variável resposta binária. A Regressão Logística atribui diferentes pesos para cada uma das variáveis explicativas do modelo que, juntas, fornecem a probabilidade de o cliente pertencer ao grupo de interesse (HOSMER; LEMESHOW, 1989). No presente estudo, a variável binária indica se o cliente apresentou atraso no pagamento de suas dívidas ou não, sendo o grupo de interesse os clientes que não atrasaram o pagamento de suas dívidas (ou seja, foram bons pagadores).

Os modelos de *Credit Scoring* incluem diversas informações dos clientes, tais como dados cadastrais e comportamentais. Dessa forma, o número de variáveis que podem ser incluídas nos modelos de crédito é amplo, sendo necessário um processo refinado de extração e seleção das variáveis mais importantes na explicação da variável resposta (MANFIO, 2007; SICSÚ, 2010).

Segundo Kahmann (2013), existem diversas técnicas de mineração de dados que realizam redução da dimensão de dados originais e/ou seleção de variáveis para utilização em modelos estatísticos. Essas técnicas incluem a análise multivariada de dados, como Análise de Componentes Principais (PCA), bem como métodos computacionais de busca sequencial de variáveis, como o método *Stepwise*.

Embora já amplamente conhecidos e utilizados, nenhum desses dois métodos de seleção de variáveis considera as medidas de desempenho práticas dos modelos na escolha das variáveis explicativas. Dessa forma, surge a necessidade de verificar se a utilização de medidas de desempenho dos modelos estatísticos (tais como o

resultado do teste de Kolmogorov-Smirnov - KS) é uma boa forma de escolha e seleção de variáveis para posterior ingresso em modelos de regressão.

A justificativa para a utilização de medidas de desempenho na seleção das variáveis é o possível ganho de explicação do modelo construído. Sendo assim, é necessário avaliar as performances das técnicas de PCA e *Stepwise*, bem como de uma nova forma de seleção de variáveis baseada no resultado do teste KS, denominada como “Método Eilert”, na predição da variável resposta do modelo. Sendo assim, o objetivo do presente estudo é comparar modelos de *Credit Scoring* utilizando diferentes métodos de seleção de variáveis: PCA, *Stepwise* e Método Eilert.

2. REFERENCIAL TEÓRICO

2.1 Conceito de risco e etapa do ciclo de crédito

Há diferentes produtos de crédito oferecidos por empresas e instituições financeiras tais como empréstimos, cartões de crédito, financiamentos e outros. Todos esses produtos são passíveis de inadimplência, isto é, existe a probabilidade do não pagamento da dívida pelo cliente, também denominado como risco de crédito. A ausência de risco de crédito é utópica, pois sem risco não há perda e também não há lucro, ou seja, não há negócio (MANFIO, 2007).

De acordo com Manfio (2017), o ciclo de crédito abrange fases e funções de risco que sucedem e se inter-relacionam. A etapa inicial consiste no planejamento do negócio, ou seja, qual(is) produto(s) de crédito será(ão) oferecido(s) ao cliente. A segunda etapa consiste na comercialização do produto de crédito, ou seja, na iniciação do cliente ao crédito. A terceira etapa consiste na manutenção de contas, cujo objetivo é manter o bom relacionamento da empresa com o cliente. A quarta e última etapa ocorre quando há problemas no relacionamento da empresa com o cliente, ou seja, quando há atrasos no pagamento das dívidas. Essa etapa denomina-se como cobrança, que é quando a empresa busca ativamente o cliente para recuperação do valor monetário concedido, visando evitar perdas financeiras.

A definição de perda é complexa, pois varia entre as empresas/instituições financeiras. Geralmente é utilizado o tempo de atraso no pagamento do produto concedido para definição da perda. Geralmente, clientes que tomaram crédito e

causaram perdas não aceitáveis pelo credor são denominados “maus pagadores”, sendo o restante dos clientes classificados como “bons pagadores”. Em algumas instituições financeiras há clientes que não utilizam produtos de crédito, sendo classificados como “clientes neutros” (SICSÚ,2010).

Há diferentes formas de avaliação do risco de crédito das operações financeiras nas empresas. As avaliações podem ser realizadas de forma subjetiva, através da experiência de analistas de crédito, ou de forma quantitativa, utilizando técnicas estatísticas que mensuram o risco de inadimplência dos clientes. Avaliações subjetivas, mesmo incorporando a experiência dos gestores/analistas, não quantificam o risco de crédito, pois não estimam as perdas/ganhos das operações financeiras. Além disso, esse tipo de avaliação não é estável, pois as conclusões são divergentes entre os analistas, podendo gerar prejuízos financeiros e morais não somente para as instituições, como também aos clientes. Sendo assim, surge a necessidade da substituição desses critérios de avaliação subjetivos pelo uso de técnicas quantitativas que melhorem a tomada de decisão das empresas, não apenas diminuindo a concessão de crédito aos maus pagadores, como também aumentando o crédito aos potenciais bons pagadores (SELAU & RIBEIRO, 2009; SICSÚ, 2010).

2.2 Modelos de *Credit Scoring*

Os modelos de análise de crédito (*Credit Scoring*) são construídos combinando diferentes características (variáveis) dos usuários e produtos financeiros, cuja principal finalidade é predizer e/ou classificar o bom e o mau pagador. As técnicas baseiam-se em identificar as variáveis que são importantes para a classificação do bom ou mau cliente e o peso de cada uma dessas variáveis na avaliação. Baseados nos dados dos clientes, são determinados os escores de risco de crédito. Geralmente convenciona-se que escores mais altos predizem possíveis bons pagadores (TOMAZELA, 2007; SELAU & RIBEIRO, 2009).

A etapa do ciclo de crédito do cliente e o tipo de informação utilizada é o que diferencia os tipos de modelos de *Credit Scoring*. Os modelos baseados predominantemente em informações cadastrais, criados geralmente para novos clientes, são denominados como *Application Scoring*. Quando a instituição já conhece seu cliente, ou seja, quando há características comportamentais das operações de

crédito e das movimentações financeiras do usuário, são criados modelos de *Behavioral Scoring*. Para os clientes da etapa final do ciclo de crédito, ou seja, clientes que apresentam algum atraso nos produtos de crédito, são construídos modelos de *Collection Scoring,* que estimam a probabilidade do pagamento da dívida estabelecida (SICSÚ, 2010).

2.3 Regressão Logística em *Credit Scoring*

Os modelos de regressão podem ser definidos como equações matemáticas que demonstram o efeito de uma ou mais variáveis explicativas em uma resposta de interesse. Dessa forma, é possível estimar valores de variáveis respostas desconhecidas com base nas informações de uma ou mais variáveis conhecidas que se relacionam com a resposta de interesse (CORRAR *et. al*., 2007; CAMARGOS *et. al*., 2012).

Nos modelos de regressão linear, o relacionamento entre as variáveis é demonstrado através da equação, em que os valores da variável resposta são quantitativos. Porém, em diversas áreas do conhecimento, a variável de interesse é categórica, sendo muitas vezes de natureza binária. Sendo assim, em muitos desses casos, não se deseja estimar os valores da resposta de interesse, mas sim a probabilidade de um indivíduo ou fenômeno pertencer a uma categoria de interesse (ARAÚJO, 2007; CORRAR *et. al*., 2007).

Diante disso, por volta da década de 60, foi desenvolvida para a área de ciências médicas uma técnica de regressão para predição de variáveis respostas binárias, onde é possível discriminar duas classes de indivíduos ou objetos de interesse, sendo denominada como Regressão Logística (RL) (MINUSSI, 2002).

De forma geral, a RL consiste na análise do efeito de diversas variáveis explicativas (categóricas ou métricas) sobre uma variável resposta de interesse, sendo bastante utilizada na área de crédito. Nesse caso, a RL utiliza a combinação do efeito de diferentes informações cadastrais e/ou comportamentais dos clientes para predição da probabilidade de inadimplência, atribuindo pesos para cada uma das variáveis explicativas do modelo que, combinadas, fornecem a probabilidade do cliente pertencer ao grupo de bons ou maus pagadores (HOSMER; LEMESHOW, 1989; MINUSSI, 2002; CORRAR *et. al*., 2007).

Outra técnica estatística que tem propósito semelhante à RL é a Análise Discriminante (AD). Porém, de acordo com Missuni (2002), a vantagem da utilização da RL ao invés da AD está na flexibilidade da primeira, pois o modelo logístico não apresenta pressuposições quanto à forma funcional das variáveis independentes. Sendo assim, possivelmente o número de parâmetros envolvidos na estimação do modelo é menor.

Segundo Camargos *et. al*. (2012), nos modelos de RL a probabilidade de ocorrência dos eventos de interesse pode ser estimada de forma direta. No caso binário, a variável resposta Y assume somente dois valores possíveis (0 ou 1). Sendo assim, com um conjunto de p variáveis explicativas ($X_1, X_2, ..., Xp$), o modelo de RL pode ser descrito da seguinte forma:

$$\mathrm{P}(\mathrm{Y}=1)=\frac{1}{1+e^{-g(x)}}$$

$$g(x)=\beta 0+\ \beta 1X1i+\cdots+\ \beta pXpi$$

onde $g(x)$ representa a combinação linear dos coeficientes e variáveis explicativas do modelo.

2.4 Métodos de seleção de variáveis

Os modelos de *Credit Scoring*, principalmente os de *Behavioral Scoring* e de *Collection Scoring* são construídos com base em uma grande quantidade de informações dos clientes. Além das informações cadastrais, também são utilizadas informações de crédito (tipo de produto, valor monetário concedido, número de parcelas estabelecidas, número de parcelas pagas, etc.), de conta corrente (saldo de crédito, saldo de débito, número de transações financeiras, etc.), de *bureau* de crédito (presença ou ausência, quantidade e valor de restritivos, etc.) e outras. Sendo assim, o número de variáveis (características) dos clientes nesse tipo de modelo é grande, pois além das informações citadas anteriormente, também podem ser utilizadas informações temporais, ou seja, valores das mesmas variáveis em diferentes momentos ao longo do tempo (MANFIO, 2007; SICSÚ, 2010).

Dessa forma, um aspecto extremamente importante na construção de modelos de *Credit Scoring* é o processo de seleção de variáveis, isto é, dentre todas as

possíveis combinações de variáveis importantes para predição do risco de crédito dos clientes, quais devem ser selecionadas. Assim como no processo de concessão de crédito, a seleção de variáveis também pode ser realizada de forma subjetiva (com o analista escolhendo quais variáveis irão ser selecionadas para ingresso no modelo de *Credit Scoring*), bem como de forma objetiva, com a utilização de métodos de seleção de variáveis ou de redução de dimensão de variáveis do banco de dados original (MANFIO, 2007; SICSÚ, 2010).

Segundo Guyson e Elisseeff (2003), há muitos benefícios na utilização de potenciais métodos de seleção de variáveis, como a consequente facilidade de visualização dos dados, a redução das dimensões dos bancos de dados para análise e o armazenamento de informações, além da melhora no desempenho das previsões.

De acordo com Kahmann (2013), as sistemáticas de seleção de variáveis podem ser realizadas tanto para seleção de variáveis para a predição (cujo objetivo é encontrar um conjunto de variáveis explicativas que melhorem a predição de uma ou mais variáveis respostas), como para seleção de variáveis de classificação, ou seja, para encontrar o conjunto de variáveis explicativas com melhor discriminação para categorização de grupos de observações. Para isso, existem diversas técnicas de mineração de dados que realizam redução da dimensão de dados originais e/ou seleção de variáveis tais como Análise de Componentes Principais (PCA) e o Método *Stepwise*.

A PCA (do inglês *Principal Component Analysis*) é uma técnica multivariada amplamente conhecida, que procura o máximo de explicação da variância dos dados através da combinação linear entre as variáveis. A ideia da PCA é, por meio da combinação linear das variáveis originais, criar novas variáveis (denominadas componentes principais) que retém o máximo possível da informação contida nas variáveis originais (NETO & MOITA, 1998; MORAIS, 2011)

Segundo Montgomery (2004), quando há diversos parâmetros para serem estudados (cujos efeitos são inter-relacionados), ou quando alguns parâmetros são parciais ou medidas de outros parâmetros, surge a necessidade da utilização de análise multivariada.

As componentes principais são combinações das variáveis originais que apresentam ausência de autocorrelação. Com isso, após a extração das componentes, pode-se utilizar modelos de regressão, que geralmente têm como pressuposto a ausência de multicolinearidade. As componentes principais são extraídas na ordem do mais até o menos explicativo. Dessa forma, as primeiras componentes da análise são as que apresentam maior explicação dos dados originais (RIGÃO, 2005; MORAIS, 2011).

Segundo Johnson & Wichern (1992), geralmente, grande parte da variabilidade total (mais de 70%) do conjunto inicial de dados podem ser explicadas por uma, duas ou três componentes principais. A variância total explicada por cada uma das componentes principais é representada pela razão da variância (autovalor) correspondente e da soma total das variâncias.

Segundo Moraes (2011), na PCA são necessárias algumas etapas para extração dos componentes principais, conforme apresentado na Figura 1.

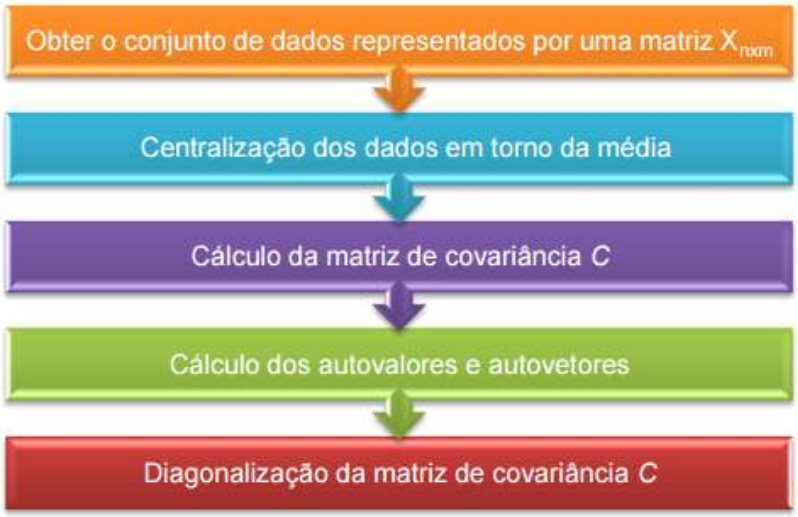

Figura 1 - Etapas da PCA
(Fonte: MORAIS, 2011)

Na PCA, não há uma única regra para definição do número de componentes finais que devem ser extraídas. Estudos mais recentes têm incluído na análise final apenas as componentes que apresentarem autovalor superior a um (MORAIS, 2011).

Já o método *Stepwise* é o mais comum dos métodos de busca sequencial, baseado na adição ou remoção de variáveis dos modelos estatísticos com base na significância de cada uma dessas variáveis. O método possibilita examinar a contribuição adicional de cada variável explicativa ao modelo, sendo bastante utilizada em técnicas estatísticas como regressão linear e regressão logística (ALVES *et. al*., 2013; ZIMMER & ANZANELLO, 2014).

O objetivo do método *Stepwise*, apresentado na Figura 2, é selecionar as variáveis que maximizam a previsão da resposta com o menor número de variáveis explicativas empregadas. Dessa forma, identifica-se as variáveis estatisticamente mais significantes para compor o modelo final. Um ponto fraco desse tipo de método de seleção de variáveis é que seu desempenho é afetado por variáveis ruidosas e correlacionadas (CORRAR *et. al*., 2007; ARTUSO & NETO, 2014).

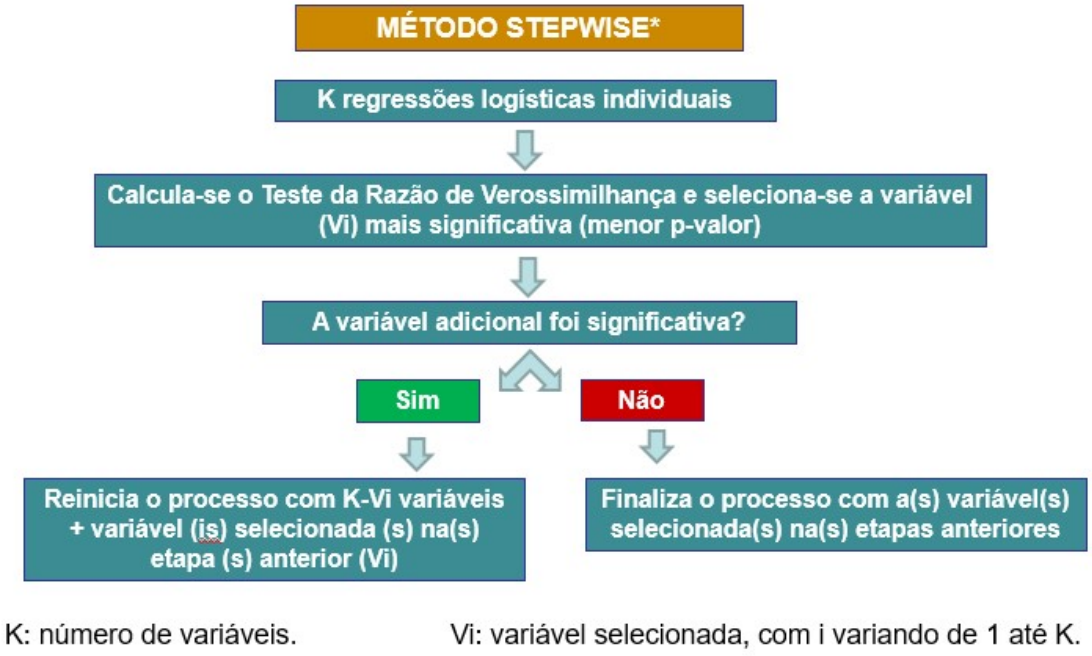

Figura 2 - Etapas do método *Stepwise* em Regressão Logística
(Fonte: Baseado em CORRAR *et. al*., 2007)

2.5 Avaliação dos modelos de *Credit Scoring*

Um aspecto fundamental em modelos de *Credit Scoring* é a avaliação do desempenho dos modelos construídos. Essa avaliação verifica se o ajuste do modelo foi realizado adequadamente e se a predição de bons e maus clientes está sendo realizada de forma satisfatória, atendendo os objetivos da instituição (TOMAZELA, 2007).

Existem diversas formas de verificação e medição do desempenho do modelo de crédito. Algumas das medidas utilizadas para esse fim são o percentual de acerto geral, os valores de sensibilidade e especificidade, o resultado do teste de Kolmogorov-Smirnov (KS), a área sob a curva de Lorentz ou ROC (*Receiving Operational Characteristic),* o coeficiente de Gini, a diferença entre as taxas de inadimplência nas faixas extremas do escore definidas pelos decis (DTI), a área entre as curvas da distribuição acumulada dos escores (AEC) e outros (TOMAZELA, 2007; ALVES, 2008; PACHECO & SELAU,2016).

Uma das medidas de avaliação dos modelos de crédito mais conhecida e utilizada é a estatística KS. Esse teste é baseado na medição da distância entre as distribuições de probabilidades dos clientes adimplentes e dos inadimplentes. O valor do teste, cujo intervalo assume valores de 0 até 100%, é obtido através do valor máximo da distância entre a função de distribuição acumulada de maus clientes e a função de distribuição acumulada de bons clientes. Esse teste é bastante utilizado uma vez que apresenta valores conhecidos de desempenho e segue como referência para comparação dos modelos de crédito construídos em diferentes empresas e instituições financeiras. Os valores de referência do KS variam conforme o modelo de crédito construído. Modelos de iniciação ao crédito geralmente apresentam valores de referência mais baixos quando comparados com modelos de análise de comportamento (*Behavioral Scoring*). Dessa forma, na Tabela 1, é apresentada uma tabela geral com valores de desempenho do KS utilizados no mercado de crédito para avaliação do desempenho de modelos de *Behavioral Scoring* construídos (TOMAZELA, 2007; ALVES, 2008, PACHECO & SELAU,2016).

Tabela 1 - Valores de Referência do KS

KS	Nível de Discriminação
<= 40%	Baixo
> 40% e <=50%	Aceitável
< 50% e <= 60%	Bom
< 60% e <= 70%	Muito bom
> 70%	Excelente (pouco usual)

(Fonte: SICSÚ, 2010)

Um aspecto importante ao realizar a avaliação das medidas de desempenho dos modelos de *Credit Scoring* é o balanceamento amostral de bons e maus clientes. O resultado do desbalanceamento amostral consiste em pior predição e redução de algumas das medidas de desempenho (como a proporção de acerto) dos modelos. Um exemplo de casos semelhantes são amostras onde há grande número de bons clientes, resultando em melhor predição desse grupo de clientes. Sendo assim, para melhorar a predição e o modelo final construído, a amostra de desenvolvimento pode conter proporções iguais de bons e maus clientes (50%). Outra forma de melhorar a predição e o desempenho final é alterar os pontos de corte dos escores de classificação dos clientes. Caso a empresa prefira classificar melhor os maus clientes, deve aumentar o ponto de corte de classificação e vice-versa (BROWN & MUES, 2012).

3. METODOLOGIA

O estudo foi estruturado com base nas etapas de criação de modelos de *Credit Scoring* (Figura 3) sugeridas por Selau e Ribeiro (2009). As informações utilizadas são de uma base de dados reais, provenientes de uma instituição financeira que concede diferente produtos de crédito aos seus clientes. Tendo em vista a confidencialidade das informações dos clientes e da empresa, todos os dados de identificação dos clientes foram omitidos e o nome das variáveis utilizadas para construção do modelo final foram substituídas por denominações genéricas.

Figura 3 - Etapas para construção de modelos de *Credit Scoring*
(Fonte: SELAU & RIBEIRO, 2009)

O modelo de *Credit Scoring* do estudo foi realizado com a utilização da técnica de Regressão Logística. Foram analisadas primeiramente 90 variáveis, construídas a partir de informações cadastrais e comportamentais dos clientes da empresa. Todas as variáveis foram categorizadas a partir do algoritmo "*Interactive Grouping*" do *Software SAS Enterprise Miner* 6.1 (SAS, 2012).

Com as variáveis categorizadas, foram realizados diferentes métodos para identificação e seleção das variáveis mais importantes para predição dos bons e maus clientes. Os métodos utilizados nessa etapa do estudo foram PCA, *Stepwise* e uma nova forma de seleção de variáveis denominada método Eilert, que foi desenvolvida e é utilizada pela instituição financeira que forneceu os dados para o presente estudo.

O método Eilert consiste em selecionar as variáveis mais importantes para o modelo de *Credit Scoring* de forma semelhante ao método *Stepwise*, porém baseado nos valores de desempenho obtidos pelo teste KS, em substituição ao nível de significância usual das variáveis adotado pelo método *Stepwise*. O algoritmo utilizado por este novo método para seleção das variáveis do modelo de *Credit Scoring* segue as etapas apresentadas na Figura 4.

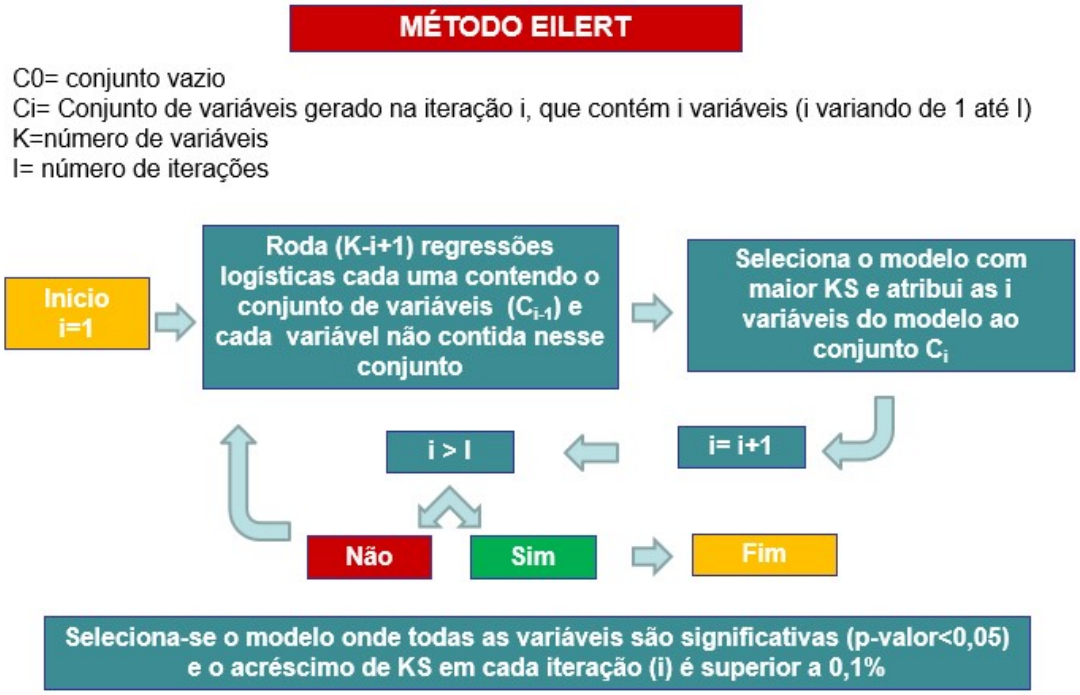

Figura 4 - Algoritmo do método Eilert

De forma simplificada, o método consistiu nas seguintes etapas:

1. Inicialização do processo, com a realização de 90 regressões logísticas (i=1) individuais (com cada uma das variáveis do banco de dados originais).
2. Avaliação do valor do KS de cada um dos 90 modelos construídos, reservando o modelo com a variável que obteve o melhor resultado da medida de desempenho (KS);
3. São geradas 89 novas regressões logísticas de duas variáveis (i=2), contendo a variável da etapa anterior e adicionando cada uma das demais variáveis. Novamente verificou-se o valor de KS gerado por cada um dos 89 modelos construídos, reservando o modelo de duas variáveis que apresentou o melhor desempenho do teste KS;

4. Nas etapas seguintes, são repetidos progressivamente os passos anteriores, até que se atinja o número de iterações estabelecido à priori, descrito na próxima etapa. A variável incluída em cada etapa sempre foi a que, em conjunto com as demais selecionadas nas etapas anteriores, apresentou o maior incremento no KS.

5. Após o processamento de todas as i iterações, ao final do processo, teremos uma tabela com i variáveis selecionadas em cada iteração, o KS cumulativo e a quantidade total de variáveis não significativas até a iteração correspondente. Dessa forma, sugere-se utilizar no modelo final todas as variáveis até a primeira das seguintes condições ser atendida: Iteração anterior ao surgimento de uma variável não significativa (p-valor estabelecido pelo pesquisador) ou até a iteração que precede três iterações consecutivas com acréscimo de KS inferior a 0,10%

A seleção de variáveis através do método de PCA consistiu em extrair os componentes principais a partir do critério de Kaiser, ou seja, todos os que obtiveram autovalores superiores a um (FLECK & BOURDEL, 1998).

No método *Stepwise*, as parametrizações adotadas foram os valores de referência usualmente utilizados para definição do nível de significância dos modelos estatísticos (0,05 e 0,01). Em um primeiro momento, utilizou-se como parametrização para entrada no modelo as variáveis com significância estatística de 5% (p-valor igual ou inferior a 0,05). Já no segundo momento, optou-se por utilizar as variáveis com significância estatística de 1% (p-valor igual ou inferior a 0,01). Dessa forma, foram avaliadas as duas parametrizações como métodos diferentes para seleção de variáveis do modelo final.

Após a aplicação das diferentes técnicas de seleção de variáveis, foi utilizada a técnica de RL para predição do resultado final do modelo. Para isso, foi utilizada como categoria de referência os bons clientes, isto é, os modelos finais predizem a probabilidade de os clientes cumprirem suas obrigações financeiras. Para cada um dos conjuntos de variáveis indicadas pelos métodos de seleção, foi realizado um modelo diferente de RL.

Posteriormente, os modelos construídos com os métodos de seleção de variáveis foram avaliados e comparados, utilizando diferentes medidas de

desempenho de modelos estatísticos tais como o percentual de acerto geral, valor do teste KS e a área sob curva ROC.

4. RESULTADOS

A base de dados da empresa era constituída de 240.000 clientes, sendo 224.681 bons pagadores e 15.319 maus pagadores. Dessa forma, foram separadas as amostras, que considerou 200.000 clientes na amostra de desenvolvimento e 40.000 clientes na amostra de validação.

Na utilização dos métodos de seleção de variáveis (PCA, *Stepwise* e método Eilert) para construção dos modelos finais com a técnica de RL foi utilizada somente a base de desenvolvimento, com os 200.000 clientes.

Na Figura 5, é possível observar o efeito do incremento do KS obtido pelo método Eilert com a adição de variáveis no modelo. Dessa forma, observamos graficamente que por volta da vigésima iteração, o acréscimo do KS estava próximo de zero. Diante disso, no presente estudo, o modelo final com esse método de seleção ficou com 22 variáveis, sendo o critério de parada o acréscimo do KS inferior a 0,10% em três etapas consecutivas.

Na realização do *Stepwise* para seleção das variáveis, ocorreu um problema de convergência na matriz dos dados a partir da etapa 17 do método. Essa ausência de convergência ocorria com a presença da variável 81 do banco de dados inviabilizando a correta análise dos resultados do método. Esse problema pode ter ocorrido diante da similaridade da variável 81 com outra variável do banco de dados, uma vez que se trata de um modelo de crédito comportamental, com variáveis que já haviam sido categorizadas anteriormente. Dessa forma, para melhor análise dos dados e interpretação correta dos resultados, optou-se por retirar a variável 81 do banco de dados para análise do método *Stepwise*. Dessa forma, o banco de dados desse método consistiu em 89 variáveis explicativas. O mesmo erro de convergência ocorreu na utilização de todas as variáveis do banco de dados na composição do modelo de *Credit Scoring*. Dessa forma, para análise do desempenho do modelo sem seleção de variáveis, também foi utilizado o banco de dados com 89 variáveis. Na utilização da

PCA e do método Eilert, esse problema não ocorreu e então o banco de dados inicial para esses métodos consistiu em 90 variáveis explicativas.

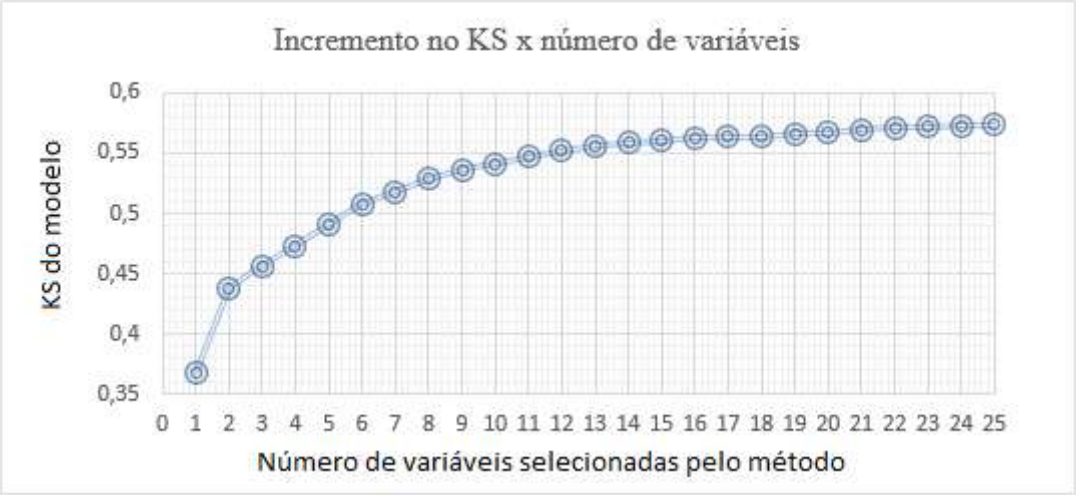

Figura 5 – Incremento do KS no método Eilert

Na Tabela 2, é possível observar o número de variáveis selecionadas nos métodos de seleção de variáveis escolhidos pelo presente estudo.

Tabela 2 – Número de variáveis selecionadas nos métodos de seleção

Seleção de variáveis	
Método	Variáveis
PCA	20
Eilert	22
Stepwise 0,05	63
Stepwise 0,01	56

Sendo assim, verifica-se que o método que selecionou o menor número de variáveis para o modelo final foi a PCA, seguido pelo método Eilert e pelo *Stepwise.* No método *Stepwise*, conforme critério de ajuste para permanência das variáveis no modelo (p-valor de 0,05 ou de 0,01), o número e as variáveis selecionadas variaram, sendo o método mais rígido (significância de 0,01) o que apresentou o menor número de variáveis selecionadas.

Na Tabela 3, pode-se visualizar as variáveis escolhidas para o modelo final de *Credit Scoring* pelos diferentes métodos de seleção. Dessa forma, as variáveis são

apresentadas, de cima para baixo, em sua ordem de relevância na explicação da variável resposta do modelo (bom e mau pagador).

Tabela 3 – Variáveis selecionadas nos métodos de seleção

Variáveis selecionadas para os modelos finais					
***Stepwise* 0,05**		***Stepwise* 0,01**		**PCA**	**Eilert**
VAR54	VAR26	VAR54	VAR26	PRI1	VAR01
VAR36	VAR55	VAR36	VAR55	PRI2	VAR50
VAR31	VAR86	VAR31	VAR86	PRI3	VAR08
VAR47	VAR45	VAR47	VAR45	PRI4	VAR86
VAR72	VAR57	VAR72	VAR57	PRI5	VAR02
VAR63	VAR11	VAR63	VAR11	PRI6	VAR31
VAR50	VAR89	VAR50	VAR89	PRI7	VAR65
VAR65	VAR16	VAR65	VAR16	PRI8	VAR62
VAR07	VAR01	VAR07	VAR01	PRI9	VAR81
VAR38	VAR58	VAR38	VAR58	PRI10	VAR63
VAR08	VAR09	VAR08	VAR09	PRI11	VAR72
VAR62	VAR23	VAR62	VAR23	PRI12	VAR07
VAR39	VAR52	VAR39	VAR52	PRI13	VAR54
VAR19	VAR17	VAR19	VAR17	PRI14	VAR47
VAR21	VAR85	VAR21	VAR85	PRI15	VAR19
VAR02	VAR59	VAR02	VAR59	PRI16	VAR13
VAR27	VAR71	VAR27	VAR71	PRI17	VAR67
VAR80	VAR13	VAR80	VAR13	PRI18	VAR38
VAR03	VAR04	VAR03	VAR04	PRI19	VAR12
VAR32	VAR43	VAR32	VAR43	PRI20	VAR21
VAR82	VAR77	VAR82	VAR77		VAR89
VAR41	VAR42	VAR41	VAR42		VAR35
VAR12	VAR46	VAR12	VAR46		
VAR51	VAR88	VAR51	VAR88		
VAR67	VAR49	VAR67			
VAR66	VAR20	VAR66			
VAR28	VAR60	VAR28			
VAR29	VAR14	VAR29			
VAR61	VAR30	VAR61			
VAR76	VAR18	VAR76			
VAR06	VAR73	VAR06			
VAR44		VAR44			

Diante disso, é possível verificar que as variáveis selecionadas pelo método Eilert e *Stepwise* podem ser inicialmente comparadas entre si, porém as mesmas não poderiam ser comparadas com a seleção do método PCA, uma vez que o mesmo utiliza combinações de variáveis para construção dos componentes. Dessa forma,

apenas para fins comparativos das escolhas dos métodos de seleção, foram utilizadas as variáveis com maior peso (coeficiente) em cada um dos componentes construídos, como se fossem as variáveis escolhidas. Para isso, foram verificadas os coeficiente de cada uma das 89 variáveis nos 20 componentes indicados pela PCA. Na sequência, para cada componente construído, foi selecionada a variável com maior coeficiente (mais importante para explicação daquele componente). Caso a variável já tivesse sido selecionada em outro componente anterior, foi escolhida a variável subsequente com maior coeficiente. Dessa forma, foi possível comparar as diferentes variáveis selecionadas pelos três métodos estudados, cujo resultado pode ser visualizado na Figura 6.

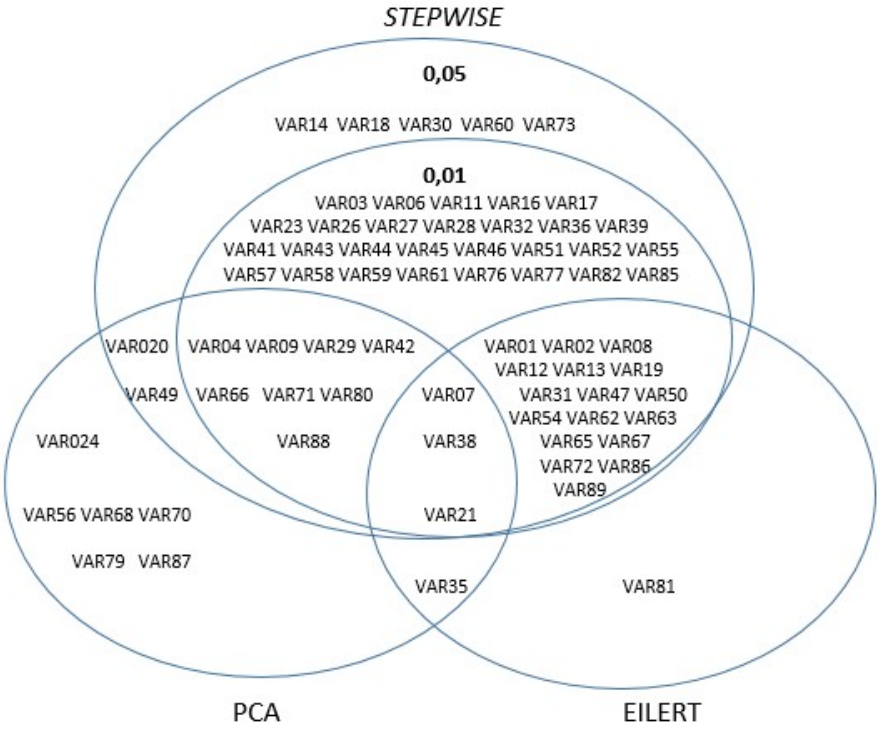

Figura 6- Diagrama dos métodos de seleção das variáveis

Através dos resultados apresentados na Figura 6, observa-se que o método Eilert e o *Stepwise* apresentaram 20 variáveis em comum, dentre as 22 possíveis (número total de variáveis do Eilert*),* ou seja, 90,9% das variáveis selecionadas no Eilert estavam contidas no método *Stepwise* (utilizando, como parâmetro de entrada de variáveis, significância tanto de 0,05 como de 0,01). Diante disso, verifica-se que as escolhas de ambos os métodos foram bastante semelhantes. Uma possível

explicação para esse resultado é uma das condições impostas pelo método Eilert para seleção das variáveis, que é o nível de significância. Isto é, variáveis que apresentem p-valor superior ao estabelecido pelo pesquisador inicialmente não são consideradas e selecionadas pelo método. Dessa forma, observa-se que, embora não realize diferentes combinações de seleções de variáveis por significância como o método *Stepwise*, os métodos Eilert e *Stepwise* apresentaram semelhanças na escolha final das variáveis do modelo.

Quando se compara as variáveis dos métodos Eilert e *Stepwise* com as variáveis de maior relevância na construção dos componentes da PCA, verifica-se que a intersecção da PCA com o *Stepwise* mais flexível (significância de 0,05) foi de 13 variáveis, ou seja, 65% de variáveis selecionadas em ambos os métodos. Na comparação entre o Método Eilert e a PCA obtive-se um total de 4 variáveis em comum, ou seja, 20% de igualdade de seleção de variáveis. Na intersecção dos três métodos, é possível visualizar três variáveis em comum entre eles, ou seja, três variáveis selecionadas independentemente do método utilizado. .

Observa-se que mesmo utilizando uma comparação subjetiva do método de componentes principais com os outros métodos, uma vez que o mesmo não utiliza as variáveis originais na construção do modelo e sim combinações delas, é possível comparar relações e semelhanças entre os métodos, verificando que existem informações importantes das variáveis que foram consideradas e utilizadas indiferente do método de seleção utilizado.

Na Tabela 4, é possível verificar algumas medidas de desempenho dos modelos construídos com as diferentes combinações de variáveis indicadas através dos métodos de seleção (PCA, Método Eilert e *Stepwise*) na amostra de desenvolvimento dos modelos.

Tabela 4 – Desempenho dos métodos de seleção na amostra de desenvolvimento

	PCA	Eilert	Stepwise 0,05	Stepwise 0,01	Sem seleção de variáveis
Nº de fatores	20	22	63	56	89
KS	56,41%	57,11%	57,98%	57,85%	58,04%
ROC	0,856	0,858	0,865	0,865	0,866
% de acerto	93,60%	93,62%	93,68%	93,68%	93,68%

Sendo assim, verifica-se que o método de seleção de variáveis que obteve o maior valor das medidas de desempenho dos modelos construídos (KS, ROC) na amostra de desenvolvimento foi o método *Stepwise* (utilizando como parâmetro de entrada a significância de 0,05). Embora o método tenha apresentado os maiores escores de desempenho dessas duas medidas, ele também foi o método que apresentou o maior número de variáveis finais (63), um acréscimo de 43 variáveis adicionais ao método de PCA e 41 variáveis adicionais ao método Eilert.

Os resultados dos modelos de RL aplicados da amostra de validação podem ser visualizados na Tabela 5.

Tabela 5 – Desempenho dos métodos de seleção na amostra de validação

Desempenho dos métodos de seleção- Amostra de Validação					
	PCA	Eilert	Stepwise 0,05	Stepwise 0,01	Todas as variáveis
Número de fatores	20	22	63	56	89
KS	56,54%	56,24%	58,06%	58,05%	58,09%
ROC	0,853	0,847	0,847	0,847	0,862
% de acerto	93,70%	93,68%	93,64%	93,71%	93,71%

Com todos os métodos de seleção de variáveis, o desempenho dos modelos foi bastante semelhante, com valores de KS em torno de 57%, ROC em torno de 0,85 e percentual de acerto por volta de 93%.

Dessa forma, verifica-se que os escores das medidas de desempenho obtidos na amostra de validação foram bem semelhantes aos obtidos no desenvolvimento do modelo, sendo o método *Stepwise* o que obteve o maior resultado do KS entre os métodos. O resultado da curva ROC foi semelhante entre os métodos Eilert e a PCA e apresentou um resultado levemente superior na utilização da PCA. O resultado do percentual de acerto também foi bastante semelhante entre todos os métodos de seleção, sendo um pouco superior no *Stepwise* com significância de 0,01.

Na base de dados do presente estudo, tanto na amostra de desenvolvimento quanto de validação, a proporção de bons era muito superior que a proporção de maus clientes. Dessa forma, o percentual de acerto não é a medida mais adequada para avaliação do desempenho final, portanto, julga-se mais importante observar nesses casos os escores obtidos através da curva ROC e do KS. Ao utilizar o método

Stepwise para a seleção de variáveis do modelo de *Credit Scoring*, observa-se um acréscimo na curva ROC e no KS final na amostra de desenvolvimento, juntamente com um elevado acréscimo do número de variáveis finais do modelo (63 variáveis com significância de 0,05 e 56 variáveis com significância de 0,01), quando comparado ao método de PCA (20 variáveis) e Eilert (22 variáveis).

Na comparação dos valores de desempenho obtidos, tanto na amostra de desenvolvimento quanto de validação, ao utilizar a PCA, obteve-se um modelo bastante semelhante em desempenho e com número bem menor de variáveis finais. Porém, a utilização desse método de seleção exige a substituição das variáveis originais por componentes “artificiais”, que são a combinação das variáveis originais. Esse método elimina eficientemente a autocorrelação entre as variáveis e melhora o desempenho do modelo, porém dificulta a interpretação dos dados e a importância e grandeza das informações contidas nas variáveis originais. Ao utilizar as variáveis originais sem métodos de seleção, obteve-se valores de ROC, % de acerto e KS um pouco superiores, tanto na amostra de desenvolvimento quanto na de validação, porém com um número bem superior de variáveis finais, o que dificulta a implementação e interpretação do modelo construído. Por tratar-se de um modelo de *Behavioral Scoring,* que inclui diversas variáveis comportamentais e históricas dos clientes, essa queda de desempenho pode ser explicada pela presença forte de autocorrelação entre diversas variáveis, efeito que acabou sendo reduzido com a utilização dos métodos de seleção.

5. DISCUSSÃO E CONCLUSÕES

O presente estudo apresentou como objetivo a comparação de três diferentes métodos de seleção de variáveis (PCA, *Stepwise* e Eilert) na construção de modelos de *Credit Scoring*.

Inicialmente, o banco de dados continha 90 variáveis históricas e comportamentais de clientes de uma instituição financeira. Porém, após a seleção de variáveis com os métodos Eilert, PCA e *Stepwise* com significância de 0,05 e 0,01, os modelos de regressão logística apresentaram respectivamente 22, 20, 63 e 56 variáveis. Diante disso, observa-se que os três métodos de seleção de variáveis foram

eficazes na redução da dimensão final de variáveis aplicadas nos modelos de *Credit Scoring* construídos no estudo.

Os resultados obtidos através das medidas de desempenho KS, ROC e percentual de acerto foram semelhantes, independentemente do método de seleção adotado para escolha das variáveis finais, apresentando escores levemente superiores ao utilizar o método *Stepwise*. Sendo assim, a definição do método adotado deve ser considerada através da facilidade de análise, interpretação e apresentação dos resultados dos modelos.

Diante disso, mesmo com escores levemente superiores de KS e ROC na amostra de desenvolvimento, o método *Stepwise* apresentou pequena redução da dimensão do banco de dados original, independentemente de utilizar critérios mais ou menos rigorosos para seleção das variáveis ao modelo final (significância de 0,05 ou 0,01). Dessa forma, na utilização desse método, constrói-se um modelo mais complexo, com elevado número de variáveis finais, mas com um pequeno acréscimo das medidas de desempenho.

Já a utilização da PCA como método de seleção de variáveis apresenta como desvantagem a dificuldade de interpretação e apresentação do modelo final, uma vez que as variáveis originais foram substituídas por componentes construídos pelo método. Dessa forma, as informações originais já não podem ser interpretadas facilmente e perde-se a facilidade de apresentação da importância das variáveis originais na predição da variável resposta do modelo.

Já o método Eilert, apesar de apresentar um número levemente superior de variáveis finais em relação ao PCA, apresentou um número bastante inferior no número de variáveis finais quando comparado ao *Stepwise*. Além disso, os escores de KS, ROC e percentual de acerto geral foram bem semelhantes aos obtidos pela PCA nas amostras de desenvolvimento e de validação, com o acréscimo de apenas duas variáveis no modelo final. Em comparação com o *Stepwise*, o método Eilert apresentou em geral escores levemente menores de KS, ROC e percentual de acerto, nas amostras de desenvolvimento e validação, porém apresentou também um número bem inferior de variáveis selecionadas ao modelo final.

Além disso, o método Eilert utilizou as variáveis originais do banco de dados, facilitando a interpretação dos resultados obtidos. Uma dificuldade da utilização desse método de seleção é a implementação computacional, pois trata-se de um procedimento com número elevado de iterações, dependendo de grande processamento dos dados. Diante disso, para uma boa utilização do método Eilert, é necessária a disponibilidade de um tempo maior de processamento e análise.

Assim como o método Eilert, o *Stepwise* apresenta como vantagem a manutenção das variáveis originais do banco de dados, facilitando a interpretação e apresentação dos resultados obtidos. Apesar de apresentar escores superiores nas medidas de desempenho, o método *Stepwise* apresentou um número bem superior de variáveis selecionadas ao modelo final. Além disso, durante o estudo, esse método apresentou problemas de convergência da matriz de dados originais, gerando resultados equivocados na análise. Em bases de dados onde há um número ainda mais elevado de variáveis e casos estudados, esse método de seleção pode não ser suficiente para redução do número original, uma vez que só usa como critério de manutenção a significância das variáveis.

Sendo assim, para uma melhor análise dos resultados e desempenho dos métodos de seleção de variáveis, julga-se necessário utilizar o mesmo número de variáveis finais indicadas por ambos os métodos. Nesse sentido, uma sugestão de estudo futuro é a comparação do desempenho de modelos de *Credit Scoring* ao utilizar diferentes combinações de variáveis indicadas por métodos de seleção.

No presente estudo, foi possível analisar as características dos três métodos de seleção das variáveis utilizadas na construção dos modelos de *Credit Scoring*. Ao comparar o método *Stepwise* e o Eilert, observa-se grande semelhança das escolhas de variáveis entre os métodos, pois um número elevado de variáveis contidas no método Eilert estavam presentes no *Stepwise*, indicando a possível importância dessas variáveis na predição do bom e mau cliente. Porém, apesar de haver discreta diferença no desempenho dos métodos de seleção de variáveis, o método *Stepwise* apresentou grande aumento no número de variáveis selecionadas em relação aos outros dois métodos. Diante disso, uma sugestão de trabalho futuro é a construção de modelos que incluam variáveis indicadas por dois ou mais métodos de seleção. Dessa forma, seria possível testar os resultados obtidos pelos métodos de seleção de

variáveis e verificar se essa combinação resulta em incremento do desempenho dos modelos.

REFERÊNCIAS

ALVES, M. C. **Estratégias para o desenvolvimento de modelos de Credit Score com inferência de rejeitados** (Dissertação de Mestrado). Universidade de São Paulo, 2008.

ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis *stepwise* aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v.1, n.1, 2013.

ARAÚJO, E. A.; CARMONA, C. U. M. Desenvolvimento de Modelos *Credit Scoring* com Abordagem de Regressão Logística para a Gestão da Inadimplência de uma Instituição de Microcrédito. **Contab. Vista & Rev.**, v.18, n.3, p.107-131, jul./set. 2007.

ARTUSO, A. R; NETO, A. C. Análise Discriminante e Regressão Logística-Reconhecimento de Padrões Para a Seleção de Portfólios no Mercado Acionário Brasileiro. **Revista da Estatística da Universidade Federal de Ouro Preto**, v.3, n.1, 2014.

BRITO, G. A. S.; NETO, A. A. Modelo de classificação de risco de crédito de empresas. **Revista Contabilidade & Finanças**, v.19, n.46, p.18-29, 2008.

CAMARGOS, M. A.; ARAÚJO, E. A. T.; CAMARGOS, M. C. S. A inadimplência em um programa de crédito de uma instituição financeira pública de Minas Gerais: uma análise utilizando regressão logística. **REGE**, v.19, n.3, p.473-492, 2012.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada: para os cursos de administração, ciências contábeis e economia.** São Paulo, Atlas, 2007.

FLECK, M. P. A; BOURDEL, M. C. Método de simulação e escolha de fatores na análise dos principais componentes**. Ver. Saúde Pública**, v.32, n.3, p.267-272, 1998.

GUYON, I.; ELISSEEFF, A**. An introduction to variable and feature selection. Journal of machine learning research**, v.3, p.1157-1182, Mar. 2003.

HOSMER, D.; LEMESHOW, S. **Applied Logistic Regression**. New York: Wiley, 3.ed., 2013.

KAHMANN, A. **Seleção de variáveis para classificação de bateladas produtivas** (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul, 2013.

MANFIO, F. **O Risco Nosso de Cada Dia.** São Paulo: Estação das Estrelas, 2007.

MINUSSI, J. A.; DAMACENA, C.; NESS J. R., W. L. Um modelo de previsão de solvência utilizando regressão logística. **Revista de Administração Contemporânea**, v.6, n.3, p.109-128, set./dez. 2002.

MORAIS, J. T. G. **Análise De Componentes Principais Integrada A Redes Neurais Artificiais Para Predição De Matéria Orgânica** (Dissertação de Mestrado), Universidade Federal da Bahia, 2011.

MONTGOMERY, D. C; RUNGER, G. C; HUBELE, N. F. **Estatística Aplicada à Engenharia.** 2 ed. Rio de Janeiro: Livros Técnicos e Científicos - LTC, 2004.

NETO, J. M. M; MOITA, G. C. Uma introdução à análise exploratória de dados multivariados. **Química Nova**, v.21, n.4, p.467-469, 1998.

PACHECO, M. N; SELAU, L.P.R. Medidas de Avaliação de Desempenho Utilizadas em Modelos de Credit Scoring. In: 22º SINAPE- Simpósio Nacional de Probabilidade e Estatística, Porto Alegre, 2016.

RIGÃO, M. H., SOUZA, A. M. Identificação de variáveis fora de controle em processos produtivos multivariados. **Revista Produção**, v.15, n.1, p.074-086, 2005.

SAS. Developing Credit Scorecards Using Credit Scoring for SAS® Enterprise MinerTM 12.1. **SAS Institute Inc**. Cary, NC, 2012. Disponível em: <https://support.sas.com/documentation/cdl/en/emcsgs/66008/PDF/default/emcsgs.pdf.> Acesso em 21 de novembro de 2016.

SELAU, L. P. R; RIBEIRO, J. L. D. Uma sistemática para construção e escolha de modelos de previsão e risco de crédito. **Gestão e Produção***,* v.16, n.3, p.398-413, 2009.

SICSÚ, A. L. **Credit Scoring: desenvolvimento, implantação, acompanhamento**. Blucher, 2010.

STEINER, M. T. A; NIEVOLA, J. C; SOMA, N. Y; SHIMIZU, T; NETO, P. J. S. Extração de regras de classificação a partir de redes neurais para auxílio à tomada de decisão na concessão de crédito bancário. **Pesquisa Operacional**, v.27, n.3, p.407-426, 2007.

TOMAZELA, S. M. O. **Avaliação de desempenho de modelos de Credit Score ajustados por análise de sobrevivência** (Dissertação de Mestrado). Universidade de São Paulo. 2007.

ZIMMER, J; ANZANELLO, M. J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Produção**, v.24, n.1, p.84-93, 2014.

ANEXO 1 - Exemplo de macro do SAS utilizada para Método Eilert

```
LIBNAME LIBRARY "caminho";


/*Indica ao SAS para exibir na LOG o valor contido em cada macro variável
no momento da substituição*/
OPTIONS NOSYMBOLGEN; /*Desativa

/*Indica ao SAS para exibir na LOG informações sobre a Macro executada */
OPTIONS MCOMPILENOTE=NONE; /*Desativa


/*Indica ao SAS para exibir na LOG o texto que foi enviado para o
compilador após a resolução da Macro */
OPTIONS NOMPRINT; /*Desativa


/*Indica ao SAS para exibir na LOG como foram executadas as macro
declarações no macro processador*/
OPTIONS NOMLOGIC; /*Desativa


/* INICIA VARIAVEL SELECAO COM VAZIO */
%LET SELECAO =
;

/* LISTA TODAS AS N VARIAVEIS A SEREM TESTADAS */
%LET TODAS =
 var01
 var02
 var03
 var04
 var05
 var06
 ...
 var090
;

DATA SELECAOVARS_EILERT; /* salva na work o banco de dados
"SELECAOVARS_EILERT" */

SET MODELO.VARIAVEIS_FINAIS; /*Vai selecionar do banco de dados
VARIÁVEIS_FINAIS da pasta MODELO " */

KEEP BM_NOVO &TODAS.; /* vai guardar as variáveis selecionadas como "todas"
e a variável resposta BM_NOVO*/

RUN;

/* Tabela onde estão localizadas as N variáveis listadas na macrovariável
"TODAS" */
%LET TABELA_ENTRADA = SELECAOVARS_EILERT;

/* TABELA ONDE SERAO SALVAS AS VARIAVEIS SELECIONADAS PELO EILERT OTIMO */
%LET TABELA_FINAL = MODELO.EILERT_FINAL;
```

```
/* NUMERO TOTAL DE VARIAVEIS A SEREM TESTADAS (QTD VARIAVEIS NA MACRO
VARIAVEL TODAS) */
%LET N = 90;

/* NUMERO DE ITERACOES QUE SERAO PROCESSADAS = NUMERO MAXIMO DE VARIAVEIS
DO MODELO */
%LET NIT = 30;


/* CRIA TABELA FINAL */
DATA &TABELA_FINAL.;RUN;

%macro EILERT_GRAPH();

/* LOOP PARA PROCESSAR NIT ITERACOES, GERANDO AO FINAL UMA TABELA OTIMIZADA
COM NIT VARIAVEIS */
%DO j = 1 %To &NIT.;

/* DEFINE PRIMEIRA VARIAVEL A SER TESTADA */
%LET VAR =%SCAN(&TODAS.,1);

/* DEFINE PRIMEIRO GRUPO DA ITERACAO COMO SENDO O GRUPO FINAL DA ITERACAO
ANTERIOR CONCATENADO COM A PRIMEIRA VARIAVEL
   DA ITERACAO  ATUAL */
%LET VAR1 =
&SELECAO.
%SCAN(&TODAS.,1)
; /* COLOCA A PRIMEIRA VARIÁVEL DE "TODAS" NA SELECAO DE VARIÁVEIS CRIANDO
A VAR1 */

DATA SELECAOVARS2;
SET &TABELA_ENTRADA.;
KEEP BM_NOVO &VAR1.;
RUN; /* SALVA A TABELA SELECAOVARS2 NA WORK PEGANDO A TABELA_ENTRADA E
COLOCANDO A VAR1 E A VAR RESPOSTA BM_NOVO */

/* RODA REGRESSAO LOGISTICA COM O PRIMEIRO GRUPO DO LOOP ATUALIZADO */

ods select none;
ods output Association = ROC;
ods output ModelANOVA = PValores;

PROC LOGISTIC DATA=SELECAOVARS2     PLOTS(ONLY)=NONE;
      CLASS &VAR1.;
      MODEL BM_NOVO (Event = '0')= &VAR1. / SELECTION=NONE LINK=LOGIT; /*
RODA A RL APENAS COM A VAR1 E A VAR RESPOSTA BM_NOVO */

      OUTPUT OUT=PREDICTIONS  PREDPROBS=INDIVIDUAL; /* SALVA A CLASSE
PREDITA E A PROBABILIDADE DO EVENTO '0' (BOM) */
RUN;

ods select all;


/* CALCULA KS DO MODELO GERADO NA REGRESSAO ANTERIOR */
PROC NPAR1WAY DATA=PREDICTIONS (Keep= IP_0 BM_NOVO) EDF NOPRINT;
      VAR IP_0;
      CLASS BM_NOVO;
      OUTPUT OUT=TEMP EDF;
RUN;
```

```
/* CRIA MACROVARIAVEL PVALOR */

PROC SQL Noprint;
      SELECT t1.ProbChiSq FORMAT=COMMA16.10 INTO :PVALOR FROM PVALORES t1
WHERE Effect = "&VAR." ;
QUIT;

/* CRIA MACROVARIAVEL QTD_NSIG */

PROC SQL Noprint ;
   SELECT (COUNT(t1.ProbChiSq)) INTO :QTD_NSIG FROM PVALORES t1 WHERE
t1.ProbChiSq > 0.05;
QUIT;


/* CRIA MACROVARIAVEL ROC */

PROC SQL Noprint;
   SELECT t1.nValue2 INTO :ROC FROM ROC t1
      WHERE t1.Label2 = 'c';
QUIT;

/* CRIA TABELA COM O NUMERO DA ITERACAO, O NOME DA VARIAVEL E O KS OBTIDO
NO PRIMEIRO GRUPO */
PROC SQL;
   CREATE TABLE KS_EILERT AS
   SELECT (1) AS Iteracao,
          ("&VAR.") as Variavel_Entrada,
          t1._D_ AS KS,
          (&ROC.) AS ROC,
              (&PVALOR.) FORMAT=COMMA16.10 AS PVALOR,
              (&QTD_NSIG.) AS QTD_NSIG
      FROM TEMP t1;
QUIT;

%DO i = 2 %To &N.;

%LET VAR =%SCAN(&TODAS.,&i.);

%LET VAR&i. =
&SELECAO.
%SCAN(&TODAS.,&i.)
;

DATA SELECAOVARS2;
SET &TABELA_ENTRADA.;
KEEP BM_NOVO &&VAR&i.;
RUN;

ods select none;
ods output Association = ROC;
ods output ModelANOVA = PValores;

PROC LOGISTIC DATA=SELECAOVARS2     PLOTS(ONLY)=NONE;
      CLASS &&VAR&i.;
      MODEL BM_NOVO (Event = '0')= &&VAR&i. / SELECTION=NONE LINK=LOGIT;

      OUTPUT OUT=PREDICTIONS  PREDPROBS=INDIVIDUAL;
RUN;

ods select all;
```

```
PROC NPAR1WAY DATA=PREDICTIONS (Keep= IP_0 BM_NOVO) EDF NOPRINT;
      VAR IP_0;
      CLASS BM_NOVO;
      OUTPUT OUT=TEMP EDF;
RUN;

/* CRIA MACROVARIAVEL PVALOR */

PROC SQL Noprint;
      SELECT t1.ProbChiSq FORMAT=COMMA16.10 INTO :PVALOR FROM PVALORES t1
WHERE Effect = "&VAR.";
QUIT;

/* CRIA MACROVARIAVEL QTD_NSIG */

PROC SQL Noprint;
   SELECT (COUNT(t1.ProbChiSq)) INTO :QTD_NSIG FROM PVALORES t1 WHERE
t1.ProbChiSq > 0.05;
QUIT;


/* CRIA MACROVARIAVEL ROC */

PROC SQL Noprint;
   SELECT t1.nValue2 INTO :ROC FROM ROC t1
      WHERE t1.Label2 = 'c';
QUIT;


PROC SQL;
   CREATE TABLE KS2_&i. AS
   SELECT (&i.) AS Iteracao,
          ("&VAR.") as Variavel_Entrada,
          t1._D_ AS KS,
              (&ROC.) AS ROC,
              (&PVALOR.) FORMAT=COMMA16.10 AS PVALOR,
              (&QTD_NSIG.) AS QTD_NSIG
      FROM TEMP t1;
QUIT;


DATA KS_EILERT;

SET KS_EILERT KS2_&i.;

RUN;


PROC DELETE DATA = KS2_&i. PREDICTIONS TEMP PVALORES ROC; RUN;

%END;


PROC SQL;
   CREATE TABLE SELECT1 AS
   SELECT t1.*
      FROM KS_EILERT t1
      ORDER BY t1.KS DESC;
QUIT;
```

```
PROC SQL;
   CREATE TABLE TODAS1 AS
   SELECT t1.Variavel_Entrada,
          t1.KS
      FROM KS_EILERT t1
      ORDER BY t1.KS;
QUIT;


PROC SQL Noprint INOBS=1;
      SELECT t1.Variavel_Entrada INTO :SELECAOT FROM SELECT1 t1;
QUIT;

%LET NTODAS = %EVAL(&N.-1);

PROC SQL Noprint INOBS=&NTODAS.;
      SELECT t1.Variavel_Entrada INTO :TODAST separated by ' ' FROM TODAS1
t1;
QUIT;

%LET SELECAO = &SELECAO. &SELECAOT.;
%LET TODAS = &TODAST.;
%LET N = %EVAL(&N.-1);


PROC SQL INOBS=1;
   CREATE TABLE SELECAO_KS AS
   SELECT t1.*
      FROM SELECT1 t1;
QUIT;

DATA &TABELA_FINAL.;
SET &TABELA_FINAL. SELECAO_KS;
RUN;

DATA &TABELA_FINAL.;
SET &TABELA_FINAL.;
WHERE KS ne .;
Drop Iteracao;
RUN;

%END;

PROC DELETE DATA = SELECT1 TODAS1 SELECAO_KS KS_EILERT; RUN;

%mend EILERT_GRAPH;

%EILERT_GRAPH()


DATA MODELO.SELECAO_EILERT_FINAL;

SET &TABELA_FINAL.;

RUN;
```

Anexo II - Exemplos de programas do SAS utilizado para cálculo da PCA e do *Stepwise*

```
/* PROGRAMA PARA CRIAÇÃO DOS COMPONENTES PRINCIPAIS ATRAVÉS DO MÉTODO PCA
*/


PROC PRINCOMP DATA = SASUSER.DADOS_ANALISE
OUT=WORK.PCA_DESENVOLVIMENTO(LABEL="Original Data and Principal Components
Scores for SASUSER.DADOS_ANALISE")
      PREFIX='PRIN'n
      SINGULAR=1E-08
      VARDEF=DF
      PLOTS(ONLY)=NONE
;
      VAR var01 var02 var03 var04 var05 var06 var07 var08 var09 var10 var11
var12 var13 var14 var15 var16 var17 var18 var19 var20 var21 var22 var23
var24 var25 var26 var27 var28 var29 var30 var31 var32 var33 var34 var35
var36 var37 var38 var39 var40 var41 var42 var43 var44 var45 var46 var47
var48 var49 var50 var51 var52 var53 var54 var55 var56 var57 var58 var59
var60 var61 var62 var63 var64 var65 var66 var67 var68 var69 var70 var71
var72 var73 var74 var75 var76 var77 var78 var79 var80 var81 var82 var83
var84 var85 var86 var87 var88 var89 var90;
      PARTIAL amostra;
RUN;


/* PROGRAMA PARA SELEÇÃO DAS VARIÁVEIS ATRAVÉS DO MÉTODO STEPWISE */

PROC LOGISTIC DATA= SASUSER.DADOS_ANALISE
            PLOTS(ONLY)=ALL
      ;
      CLASS var01 var02 var03 var04 var05 var06 var07 var08 var09 var10
var11 var12 var13 var14 var15 var16 var17 var18 var19 var20 var21 var22
var23 var24 var25 var26 var27 var28 var29 var30 var31 var32 var33 var34
var35 var36 var37 var38 var39 var40 var41 var42 var43 var44 var45 var46
var47 var48 var49 var50 var51 var52 var53 var54 var55 var56 var57 var58
var59 var60 var61 var62 var63 var64 var65 var66 var67 var68 var69 var70
var71 var72 var73 var74 var75 var76 var77 var78 var79 var80 var82 var83
var84 var85 var86 var87 var88 var89 var90;
      MODEL BM_NOVO (Event = '0')= var01 var02 var03 var04 var05 var06
var07 var08 var09 var10 var11 var12 var13 var14 var15 var16 var17 var18
var19 var20 var21 var22 var23 var24 var25 var26 var27 var28 var29 var30
var31 var32 var33 var34 var35 var36 var37 var38 var39 var40 var41 var42
var43 var44 var45 var46 var47 var48 var49 var50 var51 var52 var53 var54
var55 var56 var57 var58 var59 var60 var61 var62 var63 var64 var65 var66
var67 var68 var69 var70 var71 var72 var73 var74 var75 var76 var77 var78
var79 var80 var82 var83 var84 var85 var86 var87 var88 var89 var90      /
SELECTION=STEPWISE
SLE=0.05 /* SIGNIFICÂNCIA DETERMINADA PARA INGRESSO DA VARIÁVEL NO MODELO
*/
SLS=0.05 /* SIGNIFICÂNCIA DETERMINADA PARA PERMANÊNCIA DA VARIÁVEL NO
MODELO */

INCLUDE=0
LINK=LOGIT
      ;
```

```
      OUTPUT OUT=WORK.Predição_Dados(LABEL="Logistic regression predictions
and statistics for SASUSER.DADOS_ANALISE")
            PREDPROBS=INDIVIDUAL;
SCORE DATA= DADOS_VALIDAÇÃO; /* COMANDO PARA APLICAÇÃO DOS COEFICIENTES
PARA PREDIÇÃO DO MODELO NA AMOSTRA DE VALIDAÇÃO*/

RUN;
```