IV Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Comparação entre PCA e KPCA para validar um Modelo de ScoreCard: Uma Abordagem Empírica

Milton Luis Ribeiro Junior[1], Antonio Castelo Filho[2]
ICMC-USP

1 Introdução

Modelos de Scorecard são a espinha dorsal da gestão de risco em instituições financeiras. Eles são tipicamente construídos sobre modelos lineares (como Regressão Logística) por sua interpretabilidade e facilidade de calibração. No entanto, a premissa de linearidade nem sempre captura as complexas relações não-lineares inerentes aos dados de risco e comportamento do cliente.

A Análise de Componentes Principais (PCA) é uma técnica padrão para redução de dimensionalidade e mitigação de multicolinearidade. Embora eficaz, o PCA lida apenas com estruturas lineares. Em contraste, o Kernel PCA (KPCA) estende o PCA, utilizando a técnica de kernel trick para capturar relações não-lineares, transformando os dados em um espaço de maior dimensão (espaço de features).

O objetivo deste artigo é comparar o impacto e o desempenho de duas abordagens de redução de dimensionalidade – PCA (Linear) e KPCA (Não-Linear) – na construção de scorecards de risco.

2 Metodologia

2.1 Fundamentos do PCA e Redução de Dimensionalidade

A Análise de Componentes Principais (PCA) é uma técnica de redução de dimensionalidade que se fundamenta na decomposição de autovalores de matrizes de covariância, visando encontrar as direções de maior variância nos dados (Strang, 2019). Em essência, o PCA transforma um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de variáveis não correlacionadas, os componentes principais Boldrini.

[1]milton.ribeiro9@usp.br
[2]castelo@icmc.usp.br

2.2 O Kernel PCA e a Captura de Relações Não-Lineares

Para endereçar a restrição de linearidade do PCA, o Kernel PCA (KPCA) é empregado. O KPCA estende a técnica clássica ao utilizar o "truque do *kernel*"(*kernel trick*), permitindo que o método projete os dados de entrada em um espaço de *features* de dimensão potencialmente infinita ($\mathcal{F}$), onde o PCA é então realizado de forma linear. Conforme detalhado por Scholkopf.

2.3 Etapas de Modelagem

A base de dados utilizada neste estudo é a *German Credit Data*, obtida na plataforma Kaggle (kabure/german-credit-data-with-risk). Este *dataset* clássico é composto por 1000 instâncias, detalhando variáveis financeiras, demográficas e de comportamento de crédito de clientes. A variável de resposta (y) é o **Risco** (Risk), uma variável binária onde 0 representa "Bom Pagador"e 1 representa "Mau Pagador". Inicialmente, o conjunto de dados foi dividido em conjuntos de Treinamento (70%) e Teste (30%).

Todas as *features* foram padronizadas (=0,=1) utilizando o *StandardScaler*, procedimento essencial para algoritmos baseados em distância e variância como o PCA e o KPCA.

A redução de dimensionalidade foi aplicada em duas abordagens distintas, ambas ajustadas (*fit*) exclusivamente no conjunto de Treinamento padronizado. Primeiramente, foi aplicado o **PCA (Análise de Componentes Principais)** para determinar o número ideal de componentes (N) necessários para explicar 85%
da variância acumulada. Em seguida, o **KPCA (Kernel PCA)** foi aplicado, utilizando o mesmo número N de componentes e o **Kernel de Função de Base Radial (RBF)**, uma escolha comum para captura de não-linearidades complexas.

3 Resultado e Discussões

3.1 Poder Preditivo e Estabilidade

Os resultados da avaliação de desempenho dos dois *pipelines* de *scorecard* no conjunto de Teste estão sumarizados na Tabela 1.

Tabela 1: Comparação de Desempenho (AUC e PSI)

Modelo	**AUC**	**PSI (Treino vs. Teste)**
PCA-Scorecard (Linear)	**0.6501**	0.0792 (Excelente)
KPCA-Scorecard (RBF Kernel)	0.6412	**0.0164** (Excelente)

Em termos de poder discriminatório, o **PCA-Scorecard** demonstrou uma superioridade marginal, alcançando um AUC de 0.6501 contra 0.6412 do KPCA-Scorecard. Este resultado é crucial, pois invalida a hipótese de que a introdução de não-linearidades via Kernel RBF no espaço de *features* resultaria em um ganho preditivo suficiente para justificar a perda de interpretabilidade e o aumento da complexidade computacional do modelo. A performance superior do modelo linear sugere que as relações mais relevantes para o risco neste *dataset* são inerentemente lineares ou que as não-linearidades são ruído não generalizável.

Análise de Estabilidade (PSI)

Ambos os modelos demonstraram uma estabilidade notável, com valores de *Population Stability Index* (PSI) abaixo do limiar de 0.10, que é o ponto de corte para "Estabilidade Excelente". O KPCA, notavelmente, apresentou um PSI ligeiramente melhor (0.0164) do que o PCA (0.0792).

Embora o KPCA tenha um PSI inferior, o pequeno ganho de estabilidade não é suficiente para compensar sua perda no AUC e na interpretabilidade. Em validação de modelos, frequentemente se busca o modelo com maior AUC, desde que o PSI esteja dentro do limite aceitável. Como o PCA atende ao requisito de estabilidade (PSI < 0.10) e possui um AUC superior, ele se estabelece como a melhor escolha.

Variância Acumulada do PCA

O Gráfico 1 ilustra a contribuição de cada componente principal para a variância total dos dados de treinamento.

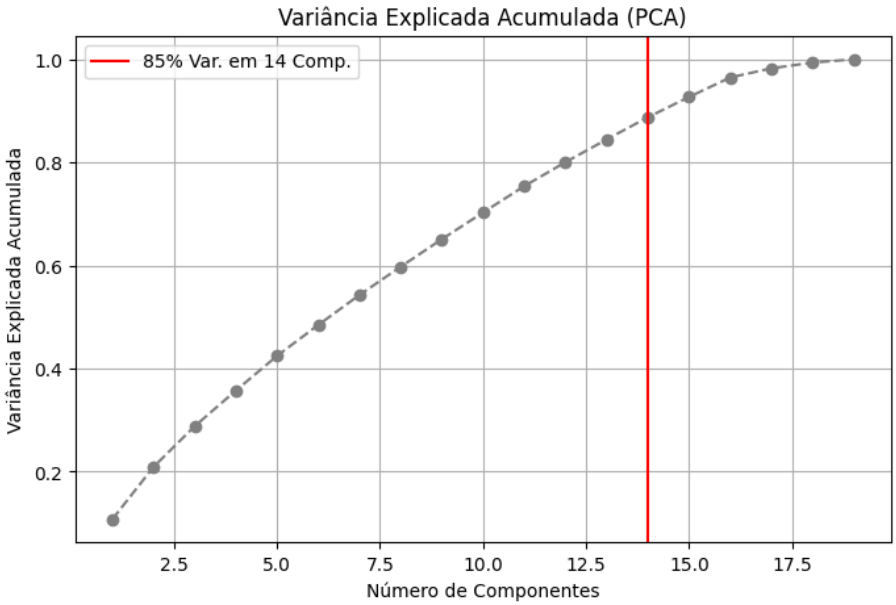

Figura 1: Variância Explicada Acumulada pelos Componentes Principais. A linha de corte indica o número de componentes retidos para atingir o limite de $\mathbf{85}\%$ da variância.

3.2 Discussão sobre a Superioridade da Abordagem Linear

Nossos resultados empíricos, onde o PCA-Scorecard demonstrou um AUC e um PSI superiores ao KPCA-Scorecard, alinham-se à cautela defendida na literatura sobre o uso de métodos complexos (*e.g.*, não-lineares) em ambientes de risco Lessmann. O PSI superior no PCA sugere que a complexidade e flexibilidade adicionadas pelo KPCA não se traduziram em maior poder de

generalização; pelo contrário, podem ter capturado ruído específico do *dataset* de treinamento, resultando em *overfitting* e, consequentemente, em menor estabilidade na validação *out-of-sample*. O resultado reforça a noção de que, para *scorecards*, a simplicidade e a robustez do método linear (PCA) frequentemente superam o ganho marginal de acurácia de métodos mais opacos e não-lineares (KPCA).

4 Conclusões

A principal descoberta empírica deste estudo é que o PCA-Scorecard superou o KPCA-Scorecard em poder preditivo (AUC) e demonstrou maior estabilidade, reforçando a validade do princípio da parcimônia (simplicidade) em ambientes de risco.

Esta descoberta sugere que as relações mais importantes para a classificação de risco no dataset [German Credit Data] são predominantemente lineares ou que os ganhos não-lineares do KPCA são superados por desvantagens práticas.

Referências

[1] J. L. Boldrini, S. I. R. Costa, V. R. Ribeiro, and H. G. Wetzler. *Álgebra Linear e Aplicações, 3a. edição*. Harbra, São Paulo, 1984.

[2] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[3] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[4] S. Lessmann, B. Baesens, H. H. K. Thomas, R. E. C. P. van den Bussche, and T. M. H. K. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124-135, 2015.

[5] G. Strang. *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press, 2019.