R e *Machine Learning* aplicado no Banco BHC

Eduardo Germano[1]

[1]Especialização em Data Science
UniRitter – Brazil

eduar.germano@gmail.com

O Banco BHC é uma importante instituição financeira no cenário nacional. A instituição possui um forte time de cientistas de dados e está passando por um processo de expansão, necessitando assim, adotar estratégias que apoiem este crescimento. Com a expansão da carteira de clientes, torna-se necessário criar políticas de crédito personalizadas para cada tipo de perfil de cliente do banco. Deste modo, nós temos a oportunidade de utilizar a linguagem R [Team 2000] e *Machine Learning* [Mitchell et al. 1997] para atender a necessidade do Banco BHC.

Entender as etapas de um ETL [Vassiliadis et al. 2002] é fundamental para o sucesso desta tarefa. Deste modo, devemos, primeiramente, agrupar as informações dos clientes em uma base e limpar os dados de forma que a mesma possa prosseguir com sua tarefa. Temos assim, a oportunidade de aplicar a técnica de *PCA* (*Principal Component Analysis*) [Maćkiewicz and Ratajczak 1993] para identificar as características mais relevantes nos dados dos clientes e reduzir a dimensionalidade do modelo de conhecimento em questão. Tal técnica tem como objetivo reduzir e otimizar o modelo de conhecimento. Entretanto, a instituição financeira definiu *a priori*, quais serão as características que devemos considerar em nossa tarefa.

Após a preparação dos dados, podemos aplicar o algoritmo *K-Means* [Hamerly and Elkan 2004], disponível como função nativa na linguagem R, para gerar a segmentação da carteira de clientes em N grupos. Tal segmentação também é conhecida como clusterização, que é o particionamento dos dados em grupos com instâncias de características similares. A utilização do *K-Means* em nosso *case* é mais adequada, pois estamos lidando com dados não-rotulados e desejamos que o algoritmo extraia o conhecimento em nosso modelo. Além disso, o *K-Means* é muito utilizado para tarefas de clusterização [Burkardt 2009].

Um ponto importante é que o time de *Data Science* poderá gerar diversas visualizações das bases utilizadas pelo *K-Means* e também por suas clusterizações. Esta etapa facilitará que um especialista do negócio analise os modelos que foram considerados e gerados, com o objetivo de entender a característica da carteira da instituição e de cada segmentação criada. Com a carteira devidamente segmentada, o Banco BHC poderá conceder linhas de crédito personalizadas, e testes A-B para entender o comportamento dos clientes de cada segmentação. Os resultados desta estratégia deverão ser bem analisadas, de modo que sejam identificadas as oportunidades para os clientes de cada grupo.

Por fim, após compreendido a característica de cada *cluster*, também poderemos treinar um algoritmo de aprendizado supervisionado para classificar novos clientes em um dos segmentos da carteira do banco. Deste modo, o Banco BHC estará preparado para

enfrentar o intenso processo de modernização de sua estratégia *data-driven*, e se manter como uma instituição financeira importante no cenário nacional.

Referências

Burkardt, J. (2009). K-means clustering. *Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics*.

Hamerly, G. and Elkan, C. (2004). Learning the k in k-means. *Advances in neural information processing systems*, 16:281–288.

Maćkiewicz, A. and Ratajczak, W. (1993). Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Mitchell, T. M. et al. (1997). Machine learning.

Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*.

Vassiliadis, P., Simitsis, A., and Skiadopoulos, S. (2002). Conceptual modeling for etl processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, pages 14–21.