UM MODELO DE REGRESSÃO LOGÍSTICA MÚLTIPLA PARA PREDIÇÃO DE *TURNOVER* – UMA APLICAÇÃO EM *PEOPLE ANALYTICS.*

Glaucia Jardim Pissinelli
gpissinelli@gmail.com

Leonardo Tomazelli Duarte
Faculdade de Ciências Aplicadas - Unicamp
R. Pedro Zaccaria, 1300, Limeira - SP
leonardo.duarte@fca.unicamp.br

Cristiano Torezzan
Faculdade de Ciências Aplicadas - Unicamp
R. Pedro Zaccaria, 1300, Limeira - SP
cristiano.torezzan@fca.unicamp.br

RESUMO

O gerenciamento eficaz dos recursos humanos de uma organização tem recebido, cada vez mais, a atenção de gestores, pesquisadores e especialistas de diversas áreas em função de sua importância crucial, não apenas para a operacionalização de atividades laborais mas, principalmente, para geração de valor nas organizações. A taxa de rotatividade do quadro de colaboradores (*turnover*) tornou-se um indicador de reconhecida relevância para as organizações, a ponto de ser utilizado como uma medida de eficiência de gestão e estabilidade organizacional. Devido ao grande volume de informação disponível, muitos métodos de análise de dados, incluíndo aprendizado de máquina e mineração de dados têm sido propostos como alternativas de apoio à decisão na área de RH. O conjunto dessas técnicas tem se tornado conhecido através do termo *People Analytics*. Neste trabalho apresentamos um modelo de regressão logística múltipla para a criação de um indicador que modela a propabilidade de cada funcionário de uma compania se desligar à pedido (turnover voluntário). O modelo foi testado com dados reais e obteve taxas de acerto superiores a 85%.

PALAVRAS CHAVE. ***People Analytics*****,** ***Turnover*****, Regressão Logística.**

Tópicos (AD&GP - PO na Administração & Gestão da Produção; EST - Estatística; OA - Outras aplicações em PO.

ABSTRACT

The efficient management of the human resources has increasingly received the attention of managers, researchers and specialists from a variety of fields. The HR has crucial importance, not only for the operationalization of labor activities, but mainly for the generation of value in the organizations. In this context, the turnover rate has become an indicator of recognized relevance to organizations, to the point of being used as a measure of organizational efficiency and management efficiency. Due to the large volume of available information, several methods of data analytics, including machine learning and data mining have been proposed as alternatives for decision support in the HR area. The set of such techniques has become known through the term People Analytics. In this paper, we present a multiple logistic regression model to derive an indicator that models the propensity of each employee leaves voluntarily a company. The model was tested with real data and high indexes of correct predictions was obtained.

KEYWORDS. People analytics. Turnover. Logistic Regression.

Paper topics (AD & GP - OR in Administration & Production Management; EST Statistics; OA - Other applications in OR.)

1. Introdução

O papel dos Recursos Humanos (RH) nas organizações tem se modificado ao longo do tempo, passando de uma função meramente administrativa para exercer um papel mais estratégico. Essa evolução ocorre pela tendência na mudança de comportamento do RH frente às exigências de um mercado cada vez mais competitivo economicamente, que exige das empresas soluções mais eficazes e de longo prazo.

Atualmente o RH possui um papel central na captação, capacitação e retenção de pessoas nas organizações. Neste contexto, a área tem absorvido e adaptado as mais variadas técnicas de gestão, incluindo o uso de modernos métodos de análise de dados. Um relatório de tendências de capital humano divulgado em 2016 revelou que 86% dos executivos de RH consideram o *people analytics* como um ponto de foco principal para as organizações (Bersin et al., 2016).

O termo *people analytics* é utilizado para designar a análise de pessoas com base em dados, através da integração de diferentes fontes de informação, dentro e fora da empresa, que procuram responder questões relacionadas aos ativos de capital humano de uma organização (Isson et al., 2016). Este tipo de análise envolve, em geral, o uso de ferramentas computacionais para mensurar, reportar e entender o desempenho dos funcionários com base em dados, resultando em análises que podem conduzir a uma melhor gestão e decisões de negócios para uma organização (Collins et al., 2017).

O capital humano de uma empresa, definido como os conhecimentos, as competências e as habilidades dos funcionários, é considerado um dos ativos mais importantes para torná-la competitiva. Qualificar e manter esse capital tem se tornado um desafio aos departamentos de RH [Fitz-enz e John R. Mattox, 2014] [Eriksen, 2016]. A retenção de bons funcionários é um problema recorrente e complexo que o RH enfrenta no dia a dia. Buscando evitar a evasão desses funcionários, as empresas tem, cada vez mais, adotado novos métodos que auxiliam na redução do *turnover*, ou taxa de rotatividade.

Tradicionalmente, o *turnover* é dividido em duas classes: voluntário e involuntário. O turnover voluntário é caracterizado por uma escolha pessoal do funcionário em deixar a empresa e buscar novas oportunidades. O turnover involuntário ocorre em casos de demissão motivada por interesses da empresa, por aposentadoria ou morte do funcionário[Chang, 2009].

A gestão do *turnover* é um desafio importante para a maioria das empresas e diversas técnicas tem sido empregadas. Mais recentemente gestores de RH estão apostando em métodos de *Big Data* (grande volume de dados) para auxiliar análises sobre a rotatividade de seus funcionário [Isson e Harriott, 2016].

Esta nova maneira de administrar pessoas contrasta com o gerenciamento clássico do capital humano na maioria das organizações, que tradicionalmente é centrado em relações interpessoais, ou tomada de decisão com base na experiência do decisor, ao invés de se utilizar das análises avançadas dos dados disponíveis (*Advanced People Analytics*). A análise de dados de pessoas oferece oportunidade para profissionais de recursos humanos posicionar-se como parceiros estratégicos da organização, adotando técnicas, mais sofisticadas, e comprovadamente eficazes no suporte a tomada de decisão.

Este tipo de análise é motivada por modelagens estatísticas preditivas e tem alcançado grande sucesso em áreas como segurança de crédito e marketing, que utilizam técnicas similares para atrair, engajar, reter e recompensar seus clientes mais valiosos. Bancos e empresas financeiras já utilizam métodos de inteligência artificial para realizar análise de crédito de clientes [Cheng e Cao, 2014] [Zhao et al., 2015] [Ozturk et al., 2016] [Li et al., 2016] [Carneiro et al., 2017] e detecção de fraudes [Van Vlasselaer et al., 2015] [Zakaryazad e Duman, 2016] [Abdallah et al., 2016]. Empresas como a Amazon e a Netflix [Isson e Harriott, 2016] usam análises preditivas para fazer recomendações de produtos a seus clientes.

Analogamente, a proposta em torno do conceito *people analytics* é utilizar métodos de inteligência artificial e modelos matemáticos para interpretar um grande conjunto de dados em busca

de respostas para conseguir atrair, engajar, crescer e reter seus principais talentos. Neste contexto, o objetivo deste trabalho é apresentar um modelo de regressão logística múltipla para a criação de um indicador que mede a intensão de turnover voluntário de cada funcionário de uma companhia. O indicador, denominado VTI (do termo em inglês *Voluntary Turnover Intention*) é totalmente baseado em dados dos funcionários, da carreira e do ambiente de trabalho da própria companhia e resulta em um número entre 0 e 100, que indica, de forma crescente a semelhança de cada indivíduo da ativa com funcionários que pediram demissão no passado. Este tipo de indicador pode ser de grande utilidade para compreender o perfil de pessoas que pedem demissão e, principalmente, para identificar casos estratégicos e propor políticas preventivas de RH e não apenas corretivas.

A metodologia proposta neste trabalho foi testada com base em dados reais de uma empresa brasileira do setor de bebidas, utilizando dados de demissões de três anos (2013, 2014 e 2015) para prever as demissões a pedido no ano de 2016 e obteve taxas de acertos superiores a 85%

O restante do trabalho está organizado da seguinte maneira: Na seção 2 apresenta-se uma breve revisão sobre regressão logística, na seção 3 descrevemos o modelo proposto e na seção 4 apresentamos alguns resultados obtidos com base em dados reais.

2. Regressão logística

A regressão logística, ou modelo logístico é uma técnica de regressão estatística utilizada para modelar variáveis dependentes que são categóricas. Casos importantes neste contexto ocorrem quando a variável dependente é binária (0 ou 1), pois tais variáveis modelam situações de falha/sucesso, perder/ganhar, morto/vivo, desligado/ligado e, no contexto deste artigo, demitido/ativo. Neste trabalho vamos nos restringir ao modelo de regressão logística binário.

O modelo logístico binário é utilizado para estimar a probabilidade de uma variável de resposta binária (dependente) ser igual a 1 com base no conhecimento de uma ou mais variáveis preditoras (ou independentes).

Formalmente, seja Y uma variável de resposta binária, onde

$$Y_i = \begin{cases} 1, \text{se a característica de interesse está presente na } i\text{-ésima observação} \\ 0, \text{caso contrário} \end{cases}$$

e $X_i = [x_{i1}, x_{i2}, \cdots, x_{ik}]$ um vetor que representa a observação de k variáveis independentes, que podem ser discretas ou contínuas. A probabilidade de $Y_i = 1$, dado que $X_i = x_{ij}$, é denotada por $\pi_i = Pr(Y_i = 1|X_i = x_{ij})$ e modelada por:

$$\pi_i = P(Y_i = 1) = f(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon),$$

onde:

- β_i são os coeficientes do modelo, que devem ser computados por métodos de máxima verossimilhança.
- x_{ij} representam os valores das variáveis independentes relacionadas à i-ésima variável dependente.
- ϵ é um ruído aditivo, que representa a parte de Y_i não explicada pelo modelo.
- f é a função logística, ou sigmoidal, dada por $f(x) = \dfrac{e^x}{1 + e^x}$.

A principal diferença entre a regressão logística e a linear é justamente a função não linear f, que mapeia o domínio x no intervalo $(0, 1)$, conforme ilustrado na Figura 1.

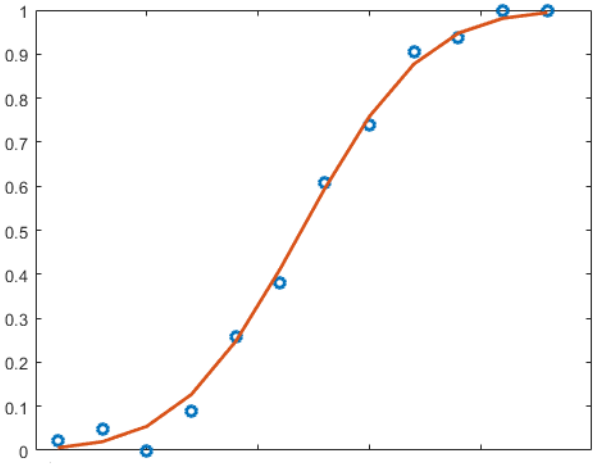

Figura 1: Gráfico da função logística (sigmoidal)

Usualmente o modelo logístico é descrito em sua forma linearizada,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon,$$

sob o qual é possível o uso de métodos clássicos de regressão para estimação dos parâmetros β_i do modelo.

Neste trabalho utilizamos a função glmfit do software MatLab calcular os coeficientes β_i e para ajustar os dados observados ao modelo proposto. Na próxima seção descrevemos como o modelo logístico múltiplo pode ser utilizado para determinar um indicador para predição de demissão voluntária em empresas.

3. Indicador de *Turnover* Voluntário

Nesta seção introduzimos um Indicador de Turnover Voluntário (ITV), que pode ser utilizado como um preditor da probabilidade de um funcionário pedir demissão em dada uma empresa. O ITV será obtido através de um modelo regressão logística multinomial, com base num conjunto de variáveis independentes, definidas à priori, e uma variável dependente binária Y, de forma que $Y_i = 1$ indica que o i-ésimo funcionário desligou-se da companhia por meio de pedido de demissão.

O problema consiste, portanto, em propor um modelo para estimar $\pi_i = P(Y_i = 1)$ para cada um dos funcionários que estão na ativa, com base em dados de funcionários que pediram demissão no passado. Em outras palavras, estimar a probabilidade do $i-$ésimo funcionário ter um perfil similar ao de funcionários que sairam voluntariamente da companhia.

3.1. Escolha e parametrização das variáveis

Para fins de facilitar a aquisição dos dados e também para auxiliar na compreensão do modelo, sugerimos que as variáveis independentes sejam alocadas em três categorias: dados pessoais, dados da carreira e dados do ambiente. A categoria de dados pessoais deve incluir variáveis típicas de descrição pessoal, como gênero, idade e estado civil, etc. A categoria dados de carreira, deve reunir variáveis que descrevam a carreira do funcionário, por exemplo, salário, tempo de companhia, nível hierárquico etc. Por último, na categoria ambiente, estão as variáveis que descrevem o ambiente de trabalho, como indice de satisfação, cumprimento de metas, avaliação pelos partes e, caso seja de interesse, até variáveis exogenas, como a situação econômica da região.

Dados pessoais	**Tipo**	**Dados de carreira**	**Tipo**	**Dados de ambiente**	**Tipo**
Gênero	Bin	Salário	Real	Satisfação com o trabalho	Int
Idade	Int	Tempo de companhia	Int	Cumprimento de metas	Bin
Estado Civil	Bin	Nível	Int	Avaliação pelos pares	Bin
Número de filhos	Int	Tempo no cargo	Int	Oferta de trabalho	Int

Tabela 1: Exemplo de variáveis independentes e respectivas parametrizações para o problema de predição de turnover voluntário.

Para que seja possível a aplicação de um modelo de regressão logística, todas as variáveis devem ser parametrizadas para formatos numéricos, preferencialmente ordinais. A tabela 1 apresenta alguns exemplos de variáveis explicativas (independentes), seus respectivos tipos e categorias.

A escolha e a parametrização das variáveis que irão compor o modelo de predição pode variar dependendo de características da companhia, por isso é fundamental a participação de gestores e representantes da equipe de RH em todas as fases do processo, principalmente na fase de modelagem.

3.2. Divisão do conjunto de dados: treinamento e teste

Como é usual em problemas de predição, o conjunto de dados disponível é dividido em duas partes: treinamento e teste. A primeira parte, *treinamento*, refere-se ao conjunto de dados que serão utilizados para o ajuste dos coeficientes do modelo. Geralmente esta parte representa a maioria do total dos dados disponíveis (tipicamente entre 60% e 80% do total). Uma parte importante dos dados é destinada para os *testes* com o modelo. Esta parte dos dados não deve ter sido utilizada na fase de treinamento e será usada para testar o grau de aderência do modelo com base em dados reais conhecidos ("prever o passado").

3.3. Ajuste do modelo

Uma vez escolhidas e parametrizadas as variáveis independentes e o conjunto de dados para o treinamento, pode-se aplicar um modelo de estimação de parâmetros para obter os coeficiêntes β_i do modelo de regressão logítico, conforme descrito na seção 2. Para os testes realizados neste trabalho, utilizamos a função glmfit do software Matlab, mas qualquer outro método de máxima verossimilhança ou pacotes estatísticos podem ser utilizados para a mesma finalidade.

3.4. Testes

Finalmente, uma vez obtidos os coeficientes do modelo que melhor se ajustam aos dados, pode-se proceder testes para verificar o grau de acerto do método. Nesta fase, para cada registro i, do conjunto de testes, calcula-se a probabilidade da variável de resposta deste registro ser igual a 1, isto é, $\pi_i = P(Y_i = 1)$.

Como a variável de interesse é binária, para verificar o grau de acerto do modelo, pode-se estabelecer um valor de referência ξ, de forma que, se $\pi_i \geq \xi$ então π_i o i-ésimo registro será associado ao valor 1 e se $\pi_i < \xi$ o registro será associado ao valor 0. Usualmente, um valor de referência adotado para o modelo logístico binário é $\xi = 0.5$ mas, alternativamente, pode ser interessante a adoção de outras linhas de corte. Por exemplo, no caso do turnover, uma companhia pode estar interessada apenas nos indivíduos que tenham $\pi \geq 0.7$. Novamente, neste ponto é fundamental o envolvimento dos gestores e do RH.

4. Resultados

Nesta seção apresentamos alguns resultados obtidos com a aplicação do método descrito nas seções anteriores para um caso real de uma empresa brasileira do setor de bebidas.

Para este problema, foram considerados dados de 8724 funcionários entre os anos 2013 à 2016. Deste total, 1304 pessoas pediram demissão, 945 foram demitidas e 6475 funcionários continuam na ativa. Ao todo foram utilizadas 19 variáveis independentes, sendo 5 na categoria de dados pessoais, 8 para dados funcionais e 6 para ambiente de trabalho. O conjunto de dados disponíveis foi dividido da seguinte forma:

- Base para treinamento: funcionários demitidos entre 2013 e 2015 + 50% da base de ativos, escolhidos aleatoriamente.
- Base para teste: funcionários demitidos em 2016 + 50% da base de ativos que não entrou no treinamento.

Considerando o valor de referência padrão, $\xi = 0.5$, a taxa de acerto do modelo foi de aproximadamente 86.3%. A curva ROC para a classificação está apresentada na Figura 2.

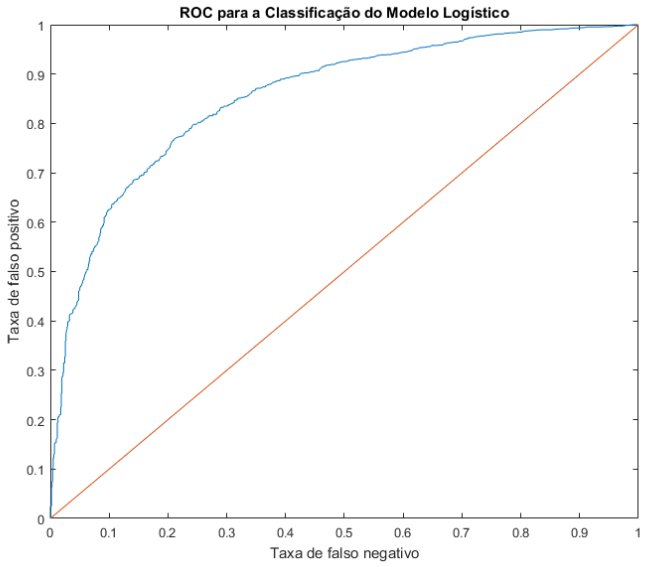

Figura 2: Curva ROC para a classificação baseada no modelo multinomial logístico.

A figura 3 mostra uma comparação entre os histogramas do indicador ITV para os funcionários que pediram demissão voluntária e para funcionários da ativa. Pode-se notar que há uma concentração maior valores altos no gráfico da esquerda e baixos no gráfico da direita. Isto significa que o indice cumpre, como esperado, um papel de classificador, indicando que funcionários com ITV altos estão associados à maior probabilidade de pedidos de demissão voluntária.

O gráfico apresentado na figura 4 mostra o percentual uma relação de compromisso entre o valor de ξ, em percentual e as taxas de voluntários/total e falso positivo. Este gráfico pode ser utilizado, na prática, como um pareto para apoio à decisão em RH, pois dependendo da disponibilidade da companhia, pode-se escolher mais ou menos funcionários para análise. Por exemplo, se a empresa desejar avaliar todos os funcionários com $ITV \times \geq 30\%$ ela estaria considerando aproximadamente 88% do total de pedidos de voluntários (curva azul), com um custo de tratar aproximadamente 30% de falsos positivos (curva vermelha).

Vale ressaltar que, para esta aplicação, o único problema em relação aos falsos positivos são os eventuais custos de campanhas preditivas em RH, visto que se um funcionário é identificado

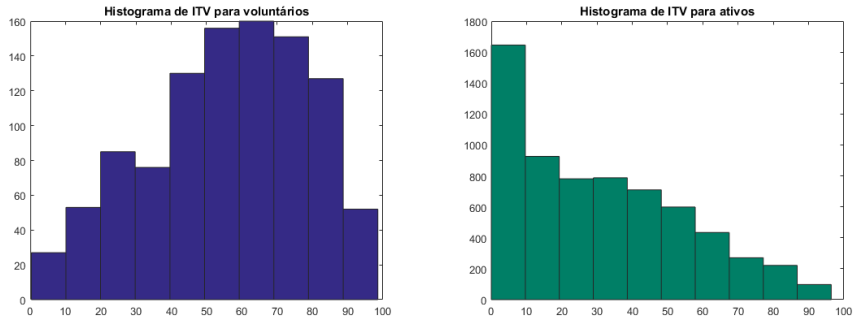

Figura 3: Comparação entre os histogramas de ITV para funcionários que pediram demissão e funcionários que estão na ativa.

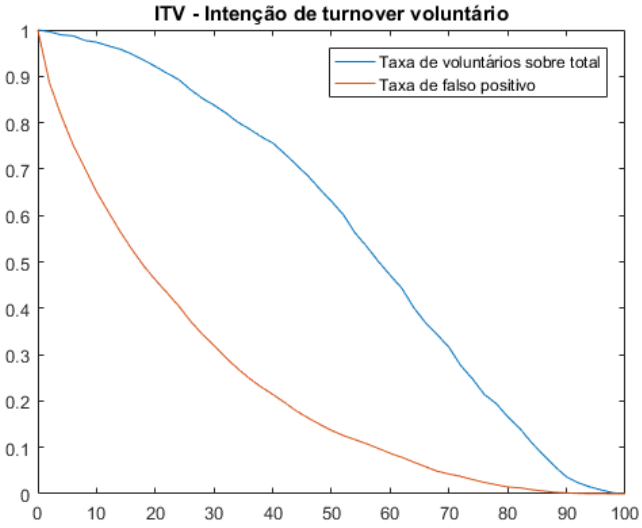

Figura 4: Relação de compromisso entre as taxas de acerto e de falso positivo do modelo em função do ITV$\times$ 100.

incorretamente como uma pessoa que tem pretenção de sair e, por isso, incluido em uma campanha para dissuadi-lo da ideia, sua intenção (real) de permanecer na companhia tende a aumentar ainda mais, não implicando em prejuízos à empresa além do tratamento de RH que for executado.

5. Conclusões

Neste trabalho apresentamos um modelo de regressão logística múltipla para a criação de um indicador que mede a intensão de turnover voluntário de cada funcionário de uma companhia. O indicador proposto é totalmente orientado a dados e foi testado com base em dados reais de uma empresa brasileira do setor de bebidas, utilizando dados de demissões de três anos (2013, 2014 e 2015) para prever as demissões a pedido no ano de 2016. Nos testes realizados o modelo obteve taxas de acertos superiores a 85%.

Este tipo de indicador pode ser bastante útil para compreender o perfil de pessoas que pedem demissão e, principalmente, para identificar casos estratégicos e propor políticas preventivas. A metodologia utilizada neste trabalho está alinhada com uma tendência moderna de gestão de RH conhecida pelo termo em inglês *People Analytics*.

Referências

Abdallah, A., Maarof, M. A., e Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113. ISSN 10848045. URL http://linkinghub.elsevier.com/retrieve/pii/S1084804516300571.

Carneiro, N., Figueira, G., e Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, 95:91–101. ISSN 01679236. URL http://linkinghub.elsevier.com/retrieve/pii/S0167923617300027.

Chang, H. Y. (2009). Employee turnover: A novel prediction solution with effective feature selection. *WSEAS Transactions on Information Science and Applications*, 6(3):417–426. ISSN 17900832.

Cheng, M.-Y. e Cao, M.-T. (2014). Evolutionary multivariate adaptive regression splines for estimating shear strength in reinforced-concrete deep beams. *Engineering Applications of Artificial Intelligence*, 28:86–96. ISSN 09521976. URL http://linkinghub.elsevier.com/retrieve/pii/S0952197613002121.

Eriksen, K. J. (2016). *The role of HR analytics in creating data-driven HRM Textual network analysis of online blogs of HR professionals*. PhD thesis, Aalto University School of Business.

Fitz-enz, J. e John R. Mattox, I. (2014). *Predictive Analytics for Human Resources*. Wiley, Hoboken, New Jersey. ISBN 978-1-118-89367-8.

Isson, J. P. e Harriott, J. S. (2016). *People Analytics in the Era of Big Data: Changing the Way You Attract, Acquire, Develop, and Retain Talent*. Wiley, Hoboken, New Jersey. ISBN 9781119050780.

Li, K., Niskanen, J., Kolehmainen, M., e Niskanen, M. (2016). Financial innovation: Credit default hybrid model for SME lending. *Expert Systems with Applications*, 61:343–355. ISSN 09574174. URL http://linkinghub.elsevier.com/retrieve/pii/S0957417416302548.

Ozturk, H., Namli, E., e Erdal, H. I. (2016). Modelling sovereign credit ratings: The accuracy of models in a heterogeneous sample. *Economic Modelling*, 54:469–478. ISSN 02649993. URL http://linkinghub.elsevier.com/retrieve/pii/S026499931600016X.

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., e Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, 75:38–48. ISSN 01679236. URL http://linkinghub.elsevier.com/retrieve/pii/S0167923615000846.

Zakaryazad, A. e Duman, E. (2016). A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing*, 175:121–131. ISSN 09252312. URL http://linkinghub.elsevier.com/retrieve/pii/S0925231215015015.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., e Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, 42(7):3508–3516. ISSN 09574174. URL http://linkinghub.elsevier.com/retrieve/pii/S0957417414007726.