LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

TRANSFERÊNCIA DE APRENDIZADO EM MODELOS DE PREDIÇÃO DE TURNOVER VOLUNTÁRIO

João Pedro de Lima Pereira
Faculdade de Ciências Aplicadas - FCA-UNICAMP
Limeira - SP
j272275@dac.unicamp.br

Cristiano Torezzan
Faculdade de Ciências Aplicadas - FCA-UNICAMP
Limeira - SP
torezzan@unicamp.br

RESUMO

Este trabalho investiga a transferência de aprendizado em métodos de predição de *turnover* voluntário. Muitas empresas, especialmente pequenas e médias, não possuem dados suficientes para treinar modelos preditivos, tornando essencial o uso de modelos pré-treinados. Investigamos o desempenho dos dois modelos mais utilizados na literatura para este problema: regressão logística (RL) e florestas aleatórias (FA). Os modelos foram treinados e testados em dois *datasets* públicos, de fontes distintas. Os resultados mostram que o classificador FA obtém acurácia superior ao RL em todos os cenários. No estudo sobre transferência de aprendizado, o modelo treinado no *dataset 1* teve boa capacidade de predizer eventos de turnover do *dataset 2*; no entanto, o segundo modelo, treinado no *dataset 2* teve um desempenho muito ruim quando testado no *dataset 1*. Isso sugere que a transferência de aprendizado pode ser dependente do contexto de treinamento, indicando a necessidade de diversidade nos dados utilizados para treinar o modelo.

PALAVRAS CHAVE. Turnover, Predição, Transferência de aprendizado.

Tópicos: BDA – Big Data e Analytics, IC – Inteligência Computacional, OA – Outras aplicações em PO.

ABSTRACT

This work investigates transfer learning in voluntary *turnover* prediction methods. Many companies, especially small and medium-sized ones, do not have enough data to train predictive models, making the use of pre-trained models essential. We investigated the performance of the two most used models in the literature for this problem: logistic regression (RL) and exploited forests (FA). The models were trained and tested on two public *datasets*, from different sources. The results show that the FA classifier outperforms the RL in all scenarios. In the study on transfer learning, the model trained on *dataset 1* had a good ability to predict turnover events in *dataset 2*; however, the second model, trained on *dataset 2* performed very poorly when tested on *dataset 1*. This suggests that transfer learning may be dependent on the training context, due to the need for diversity in the data used to train the model.

KEYWORDS. Turnover, Forecast, Transfer Learning.

Paper topics: BDA – Big Data e Analytics, IC – Computational Intelligence, OA – Other applications of OR.

https://proceedings.science/p/193706?lang=pt-br

1. Introdução

A predição de turnover voluntário, ou *attrition*, é uma área de crescente interesse na gestão de recursos humanos, pois permite que as empresas identifiquem e mitiguem a perda de talentos valiosos. Tradicionalmente, a construção de modelos preditivos eficazes requer grandes volumes de dados para treinamento. No entanto, muitas empresas, especialmente de menor porte, não dispõem de datasets extensivos, limitando a aplicabilidade dos modelos de aprendizado de máquina (ML).

Neste contexto, a transferência de aprendizado (*transfer learning*) surge como uma solução promissora. A capacidade de um modelo pré-treinado em um dataset aplicar seu conhecimento a um novo dataset pode reduzir significativamente a necessidade de grandes volumes de dados específicos da empresa. Este trabalho foca na avaliação da eficácia de dois métodos amplamente utilizados na literatura para a predição de turnover: regressão logística (RL) e Florestas Aleatórias (FA) (Akasheh et al. [2024]), explorando sua capacidade de transferência de aprendizado. Este estudo contribui para a aplicação prática de técnicas de aprendizado de máquina em cenários onde os dados são escassos, oferecendo uma solução eficaz para empresas que buscam otimizar a retenção de talentos.

2. Modelos de aprendizado de máquina para predição de *turnover*

2.1. Regressão Logística

A regressão logística é uma técnica estatística utilizada para modelar a probabilidade de uma variável dependente binária. Em vez de prever um valor contínuo, como na regressão linear, a regressão logística estima a probabilidade de ocorrência de um evento, classificando as observações em uma de duas categorias. Isso é feito ao transformar a variável dependente através da função logística, também conhecida como função sigmoide, que restringe os valores preditos ao intervalo entre 0 e 1. A equação fundamental da regressão logística é dada por:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}},$$

onde $P(Y = 1|X)$ é a probabilidade de ocorrência do evento de interesse, β_0 é o intercepto, $\beta_1, \beta_2, \ldots, \beta_n$ são os coeficientes das variáveis independentes $X_1, X_2, \ldots, X_n$. Essa abordagem tem sido amplamente aplicada em diversas áreas, incluindo a predição de turnover Zhao et al. [2019].

2.2. Florestas aleatórias

O método de Florestas Aleatórias (*Random Forest*) é uma técnica de aprendizado de máquina baseada em conjuntos, que utiliza múltiplas árvores de decisão para melhorar a precisão e a robustez das previsões (Breiman [2001]). O método funciona criando um grande número de árvores de decisão a partir de amostras diferentes dos dados de treinamento e utilizando a média das previsões dessas árvores (no caso de regressão) ou a maioria das votações (no caso de classificação) para gerar a previsão final. Cada árvore é construída utilizando uma amostra aleatória com reposição (bootstrap sampling) dos dados de treinamento e, durante a construção de cada árvore, apenas um subconjunto aleatório de características é considerado para a divisão em cada nó. Isso introduz variabilidade entre as árvores, tornando o modelo mais robusto a *overfitting*, aumentando sua capacidade de generalização.

3. Metodologia

3.1. Dados utilizados

Para realização dos experimentos numéricos deste trabalho foram escolhidos dois *datasets* públicos sobre *turnover* voluntário. O *dataset 1* (Stewart [2018]) possui dados de $N_1 = 9540$ funcionários (linhas) e 10 colunas, sendo 9 colunas atributos e 1 coluna rótulo, enquanto o *dataset 2* (Huseyn [2017]) possui $N_2 = 14999$ amostras e 10 colunas, também contendo 9 colunas atributos e 1 coluna rótulo. A escolha destes datasets levou em consideração a quantidade de atributos (colunas) similares que compartilhavam. Os atributos em comum nos dois datasets estão descritos na Tabela 1.

https://proceedings.science/p/193706?lang=pt-br

Tabela 1: Atributos considerados neste estudo

Atributo	**Descrição**
Última avaliação	Nota normalizada entre 0 e 1.
Número de projetos	A quantidade de projetos que o funcionário faz parte.
Salário	O nível salarial recebido pelo funcionário (baixo, médio ou alto).
Tempo de trabalho	A quantidade de anos que o funcionário está na empresa.
Satisfação	Satisfação do funcionário obtida através de uma pesquisa (entre 0 e 1).
Horas mensais	Quantidade de horas (em média) trabalhadas no mês pelo funcionário.
Demissão	Se o funcionário pediu demissão (igual a 1 se sim, 0 caso contrário).

3.2. Experimentos numéricos

Cada modelo foi inicialmente treinado, testado e avaliado em cada um dos *datasets* de forma independente. Para isso, cada *dataset* foi dividido na proporção de 80% para treinamento e 20% para teste. Em seguida treinamos um modelo de RL e FA para cada *dataset* e testamos o desempenho de cada modelo no *dataset* diferente do utilizado no treinamento para investigar a trasnferência de aprendizado. Os modelos treinados com o *dataset* 1 terão o sufixo ”1”e os modelos treinados com o *dataset* 2 terão o sufixo ”2”. Finalmente, utilizamos os modelos treinados em apenas um *dataset* e acrescentamos algumas linhas do outro *dataset* para entender a mudança de resultados em teste no próprio *dataset* aumentado, com um *holdout* de 20%, e também para analisar as possíveis mudanças no desempenho do modelo no *dataset* diferente ao do treinamento (*transfer learning*). A lista abaixo apresenta a denominação e uma breve descrição dos modelos utilizados:

Modelo FA-1-80%: Floresta aleatória treinada com 80% do *dataset 1*.
Modelo FA-2-80%: Floresta aleatória treinada com 80% do *dataset 2*.
Modelo FA-1-100%: Floresta aleatória treinada com 100% do *dataset 1*.
Modelo FA-2-100%: Floresta aleatória treinada com 100% do *dataset 2*.
Modelo RL-1-80%: Regressão Logística treinada com 80% do *dataset 1*.
Modelo RL-2-80%: Regressão Logística treinada com 80% do *dataset 2*.
Modelo RL-1-100%: Regressão Logística treinada com 100% do *dataset 1*.
Modelo RL-2-100%: Regressão Logística treinada com 100% do *dataset 2*.

Para avaliar o desempenho dos modelos treinados, utilizamos quatro métricas principais: acurácia, precisão, recall e F1 score. A acurácia mede a proporção de previsões corretas sobre o total de previsões, oferecendo uma visão geral da performance. A precisão avalia a qualidade das previsões positivas, indicando a proporção de verdadeiros positivos entre todas as predições positivas. O recall verifica a capacidade do modelo de identificar corretamente as instâncias positivas. Por fim, o F1 score, que combina precisão e recall através de sua média harmônica, fornece uma métrica balanceada. Essas métricas permitem uma análise detalhada da eficácia dos modelos, tanto em seu *dataset* de origem quanto na avaliação de transferência de aprendizado.

4. Resultados

Os modelos descritos na seção anterior foram implementados em linguagem Python e executados uma média de 50 vezes para cada experimento. A média das acurácias obtidas nos respectivos cenários de testes está apresentada na Tabela 2. Pode-se notar que o classificador FA supera a RL em todos os cenários, o que confirma resultados prévios da literatura (Akasheh et al. [2024]). Em relação a transferência de aprendizado, o modelo treinado no *dataset 1* teve boa capacidade de predizer eventos de turnover do *dataset 2*; no entanto, o segundo modelo, treinado no *dataset 2* teve um desempenho ruim quando testado no *dataset 1*.

Tabela 2: Acurácia dos modelos nos seus respectivos conjuntos de testes

	Dados utilizados para teste			
	Dataset - 1 (20%)	*Dataset - 2 (20%)*	*Dataset 1 (100%)*	*Dataset 2 (100%)*
Modelos	**Acurácia**			
FA-1-80%	0,857	-	-	-
FA-2-80%	-	0,990	-	-
FA-1-100%	-	-	-	0,676
FA-2-100%	-	-	0,705	-
RL-1-80%	0,721	-	-	-
RL-2-80%	-	0,767	-	-
RL-1-100%	-	-	-	0,570
RL-2-100%	-	-	0,574	-

Finalmente, visando melhorar o desempenho da transferência de aprendizado, amostras de um dataset foram inseridas no outro para efeitos de treinamento. A Figura 1 apresenta as métricas de desempenho do método FA para ambos *datasets*.

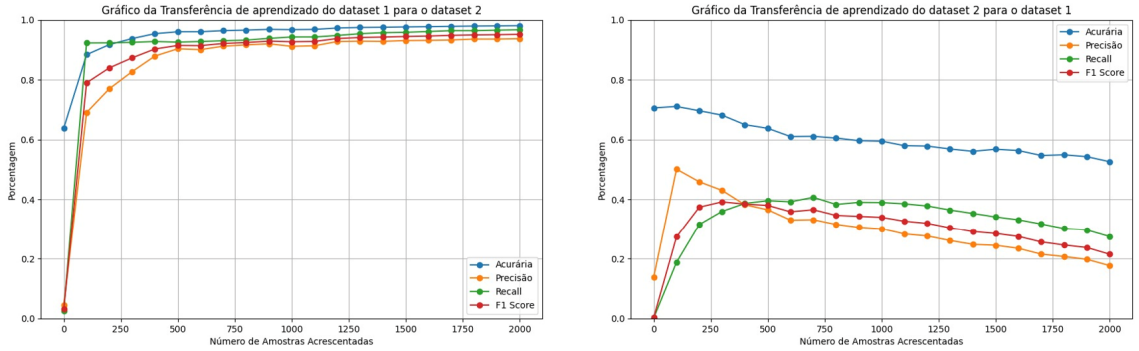


Figura 1: Efeito da introdução de novas amostras nas métricas de transferência de aprendizado.

5. Conclusão

Este estudo comparou o desempenho de regressão logística (RL) e florestas aleatórias (FA) na predição de turnover, com base em dois datasets públicos. Os resultados indicam que o modelo de FA supera a RL em todos os cenários avaliados. Além disso, na análise de transferência de aprendizado, o modelo treinado no dataset 1 demonstrou habilidade satisfatória na previsão de eventos de turnover no dataset 2, enquanto o modelo treinado no dataset 2 teve um desempenho significativamente inferior ao ser testado no dataset 1. Esses resultados destacam a importância do contexto específico de treinamento na transferência de aprendizado, enfatizando a necessidade de diversidade nos dados para um treinamento eficaz do modelo.

Referências

Akasheh, M. A., Malik, E. F., Hujran, O., e Zaki, N. (2024). A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review. *Expert Systems with Applications*, 238:121794. ISSN 0957-4174. URL https://www.sciencedirect.com/science/article/pii/S0957417423022960.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. ISSN 1573-0565. URL https://doi.org/10.1023/A:1010933404324.

Huseyn, R. (2017). Hr analytics data set. URL https://www.kaggle.com/datasets/raminhuseyn/hr-analytics-data-set/data. Accessed: 2024-06-12.

Stewart, M. (2018). Employee turnover data set. URL https://www.kaggle.com/datasets/marikastewart/employee-turnover/data. Accessed: 2024-06-12.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., e Zhu, X. (2019). Employee turnover prediction with machine learning: A reliable approach. In Arai, K., Kapoor, S., e Bhatia, R., editors, *Intelligent Systems and Applications*, p. 737–758, Cham. Springer International Publishing. ISBN 978-3-030-01057-7.