



Métodos de Aprendizado de Máquina com Preceitos Éticos para Predição de Turnover

Palavras-Chave: APRENDIZADO DE MÁQUINA, FAIRNESS, TURNOVER

Autores(as):

MARTINA LISBOA NAVARRO, LIMEIRA – UNICAMP

Prof. Dr. LEONARDO TOMAZELI DUARTE (orientador, LIMEIRA – UNICAMP)

INTRODUÇÃO:

O uso de modelos de aprendizado de máquina para previsão de rotatividade de funcionários (turnover) tem ganhado destaque devido ao potencial de apoiar decisões estratégicas em gestão de pessoas. No entanto, além da acurácia, cresce a preocupação com a imparcialidade (fairness), garantindo que previsões não favoreçam ou prejudiquem grupos específicos.

Nos últimos anos, legislações e diretrizes de governança em inteligência artificial têm enfatizado a necessidade de sistemas transparentes, justos e auditáveis. Assim, a aplicação de métricas de fairness torna-se essencial para prevenir discriminação algorítmica e assegurar que decisões automatizadas estejam alinhadas a princípios éticos. Neste contexto, este trabalho se diferencia por unir a implementação técnica de modelos preditivos à análise crítica de seus vieses, permitindo não apenas avaliar desempenho, mas também compreender e mitigar possíveis impactos sociais.

Este estudo teve como objetivo desenvolver e avaliar modelos de classificação para previsão de turnover, incorporando métricas de equidade no processo de análise. O estudo comparou o desempenho e a imparcialidade de dois algoritmos — Regressão Logística e Redes Neurais — aplicados a uma base de dados pública, considerando a perspectiva de gênero e idade como atributos sensíveis.

METODOLOGIA:

O desenvolvimento do trabalho foi estruturado em três grandes etapas interdependentes: **estudo teórico, implementação prática (base de dados e modelos utilizados) e análise de resultados (avaliação de desempenho e fairness)**.

1. Estudo Teórico

Inicialmente, realizou-se uma revisão bibliográfica abrangente sobre métodos de aprendizado de máquina supervisionado voltados para problemas de classificação, com ênfase em algoritmos amplamente utilizados na literatura, como *Decision Trees* (DT), *Random Forests* (RF), *Gradient Boosting Trees* (GBT), *Extreme Gradient Boosting* (XGB) e *Neural Networks* (NN). Para aprofundar a compreensão, foram estudadas as vantagens, limitações e aplicações ideais de cada técnica, de acordo com o contexto de uso. Paralelamente, introduziu-se o estudo das métricas de **fairness** (imparcialidade),

com base nas definições propostas por Makhoul, Zhioua e Palamidessi (2021), que estabelecem medidas quantitativas para avaliar a justiça de modelos preditivos. Essas métricas foram relacionadas à matriz de confusão, instrumento central na análise de desempenho.

2. Base de Dados

Para a aplicação prática, selecionou-se uma base de dados pública do Kaggle contendo 16 variáveis preditoras de diferentes naturezas: demográficas (idade, gênero), profissionais (tempo de serviço, setor, profissão, presença de coach, tipo de remuneração), comportamentais (meio de transporte, pipeline de contratação) e psicológicas (scores de extroversão, independência, autocontrole, ansiedade e inovação). A variável-alvo (Event) indica se o funcionário pediu demissão (valor 1) ou permaneceu na empresa (valor 0). O pré-processamento incluiu tratamento de valores ausentes, codificação de variáveis categóricas por one-hot encoding e normalização de variáveis numéricas para otimizar a performance dos algoritmos.

3. Modelos de Machine-Learning

Dois modelos de classificação foram escolhidos para implementação e comparação:

- **Regressão Logística (LR):** técnica estatística que estima a probabilidade de ocorrência de um evento a partir de uma combinação linear das variáveis independentes, com função de ativação sigmoide. É um modelo de fácil interpretação e baixo custo computacional, mas pode ser limitado em casos de alta dimensionalidade ou dados não linearmente separáveis.
- **Redes Neurais (NN):** arquitetura composta por uma camada de entrada, duas camadas ocultas (50 e 30 neurônios, respectivamente) e uma camada de saída. Utilizou-se a função de ativação ReLU nas camadas ocultas e sigmoide na saída, otimizador Adam e processo de retropropagação para ajuste dos pesos.

4. Avaliação do Desempenho

Table 1
Metrics based on confusion matrix.

	Actual positive $Y = 1$	Actual negative $Y = 0$
Predicted positive $\hat{Y} = 1$	TP (True Positive)	FP (False Positive) <i>Type I error</i>
Predicted negative $\hat{Y} = 0$	FN (False Negative) <i>Type II error</i>	TN (True Negative)

Figura 1. Métricas associadas à matriz de confusão. Fonte: Makhoul et al (2021)

O desempenho dos modelos foi medido por meio de métricas derivadas da matriz de confusão e da análise da curva AUC-ROC, que avalia a capacidade discriminativa entre as classes.

A matriz de confusão consiste em uma tabela que compara as previsões do modelo com os valores verdadeiros dos dados e é composta

por quatro elementos principais: Verdadeiros Positivos (TP - True Positives), que representa o número de casos em que o modelo previu corretamente a classe positiva; Falsos Positivos (FP - False Positives), que representa o número de casos em que o modelo previu a classe positiva, mas o verdadeiro valor era negativo; Verdadeiros Negativos (TN - True Negatives), que representa o número de casos em que o modelo previu corretamente a classe negativa; Falsos Negativos (FN - False Negatives), que

representa o número de casos em que o modelo previu a classe negativa, mas o verdadeiro valor era positivo.

A curva ROC (Curva de Característica de Operação do Receptor) é uma métrica usada para avaliar o desempenho de modelos de classificação, representando a relação entre a Taxa de Verdadeiros Positivos (TPR) e a Taxa de Falsos Positivos (FPR). A AUC (Área Sob a Curva) é a medida da área sob essa curva e indica a capacidade do modelo em distinguir corretamente as classes. No projeto, um AUC maior reflete um modelo mais eficaz em prever quais funcionários provavelmente irão pedir demissão. As matrizes de confusão e curvas ROC foram geradas para os conjuntos de treino e teste, permitindo verificar eventuais problemas de overfitting.

5. Avaliação de Fairness dos Modelos

A etapa final envolveu a análise de equidade entre grupos definidos por gênero e faixa etária, considerando como grupo privilegiado homens com mais de 25 anos (old_male). As métricas aplicadas foram:

- Equal Opportunity – igualdade da taxa de verdadeiros positivos (TPR) entre grupos;
- Predictive Equality – igualdade da taxa de falsos positivos (FPR) entre grupos;
- Predictive Parity – igualdade do valor preditivo positivo (PPV) entre grupos;
- Accuracy Equality – igualdade da acurácia geral entre grupos;
- Statistical Parity (STP) – igualdade da taxa de aceitação prevista entre grupos.

O cálculo e a visualização dessas métricas foram realizados com a biblioteca Dalex para Python, que permite diagnosticar e comparar o desempenho de modelos sob diferentes aspectos de equidade, além de identificar potenciais fontes de vies.

RESULTADOS E DISCUSSÃO:

Uma vez implementados os algoritmos de classificação de Regressão Logística e Redes Neurais, avaliou-se o desempenho através da matriz de confusão e a curva ROC. Em seguida, foram aplicadas as métricas de imparcialidade.

Desempenho Preditivo

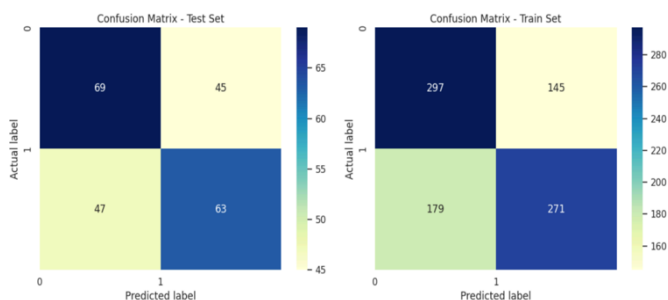


Figura 2. Matrizes de Confusão | Regressão Logística.

Mediante à avaliação dos dados obtidos nas matrizes de confusão de ambos os classificadores, destaca-se a superioridade de desempenho das Redes Neurais. Isso porque o intuito dos classificadores é maximizar o número de previsões corretas, dadas pelas variáveis de TP e TN (diagonal esquerda), bem como minimizar erros de previsões, dados pelas

variáveis de FP e FN (diagonal direita). Sob essa perspectiva, nota-se que o modelo de Redes Neurais apresentou um número maior de TP e TN para ambos os conjuntos de dados (treinamento e teste) e,

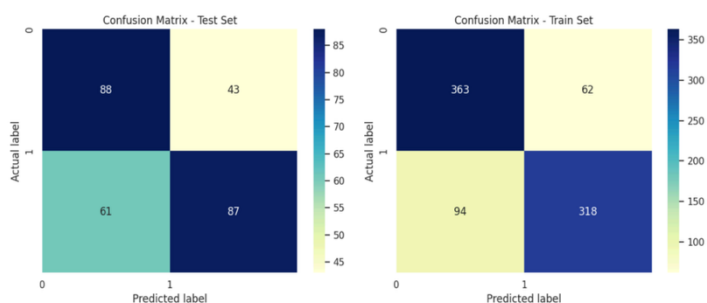


Figura 3. Matrizes de Confusão | Redes Neurais

algoritmo de Redes Neurais.

Em adição, torna-se possível avaliar as curvas ROC e suas áreas para entender o desempenho de

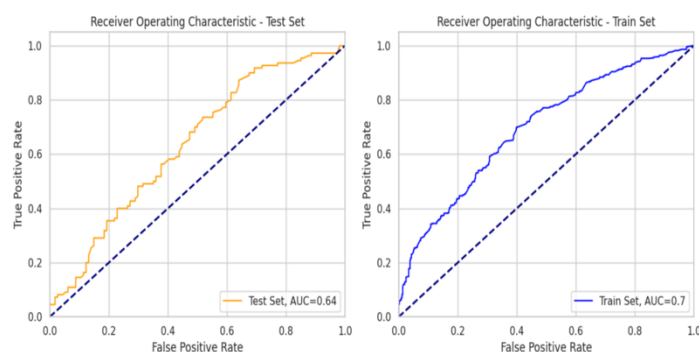


Figura 4. Curva ROC-AUC | Regressão Logística.

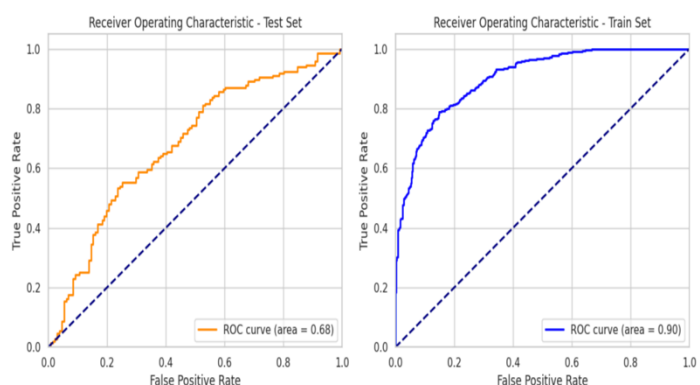


Figura 5. Curva ROC-AUC | Redes Neurais

Tabela 1. Acurácia, Precisão e Recall

	Regressão Logística	Redes Neurais
Acurácia	Treinamento: 63,67% Teste: 58,92%	Treinamento: 81,36% Teste: 62,72%
Precisão	Treinamento: 65,14% Teste: 58,33%	Treinamento: 83,68% Teste: 66,92%
Recall	Treinamento: 60,22% Teste: 57,27%	Treinamento: 77,18% Teste: 58,78%

da qual as diferenças entre os grupos são consideradas aceitáveis. Se a diferença entre as métricas de dois grupos ultrapassa esse intervalo, o modelo pode ser considerado injusto. A análise de fairness mostra que a regressão logística tem problemas mais graves, com baixa precisão para *old_female* e

portanto, maior acurácia e precisão geral, conforme as Figuras 2-3 e a Tabela 1. De maneira similar, observa-se que o Recall, que mede a proporção de verdadeiros positivos (TP) que foram corretamente identificados pelo modelo em relação ao total de instâncias que realmente pertencem à classe positiva, foi superior para ambos os conjuntos de dados no

cada classificador. Pelas Figuras 4 e 5 pode-se observar que o modelo de Redes Neurais apresentou valores superiores ao de Regressão Linear para ambos os conjuntos de dados, embora a performance para o conjunto de teste tenha sido similar, mas ligeiramente superior. Esse comportamento pode indicar um problema de overfitting, que ocorre quando um modelo aprende detalhes e ruídos específicos de um conjunto de dados de treinamento a ponto de comprometer sua capacidade de generalizar para novos dados. Isso pode ter ocorrido devido à limitação da base de dados: aumentar o conjunto de dados contribui para aumentar a generalização do modelo para o conjunto de teste.

- Análise de Imparcialidade

Em relação às métricas de imparcialidade aplicadas, observa-se que o modelo de Regressão Logística é considerado mais injusto, uma vez que 2 ou mais critérios excederam os limites aceitáveis definidos por epsilon (0.8 a 1.25). O parâmetro representa um limiar aceitável de variação entre as métricas de dois grupos distintos. Nesse contexto, ele define a faixa dentro

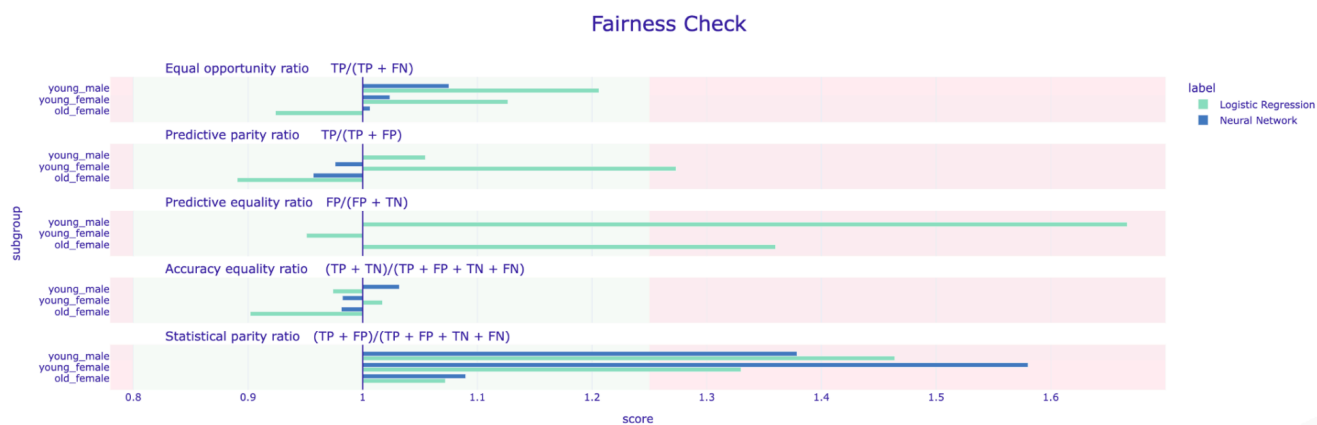


Figura 6. Comparação Métricas Fairness.

alta taxa de falsos positivos em *old_female* e *young_male*. A métrica STP indica viés significativo para *Young_female* e *young_male*, gerando disparidades nas previsões positivas. Esses fatores comprometem a justiça do modelo e a confiança dos usuários. Já o modelo de Redes Neurais apresenta melhor desempenho em fairness, embora não seja totalmente justo. A presença de valores nulos na métrica FPR aponta para dificuldades causadas pela distribuição desigual dos dados.

CONCLUSÕES:

O estudo demonstrou que a Rede Neural obteve maior acurácia e capacidade discriminativa que a Regressão Logística, mas ambas apresentaram vieses. A regressão logística mostrou desigualdades mais severas entre grupos, enquanto a rede neural concentrou seu viés principalmente na métrica de Statistical Parity. Para trabalhos futuros, recomenda-se balancear classes e garantir distribuição equitativa entre grupos, além de aplicar técnicas de mitigação de viés como reamostragem e ajuste de limiares. Também é sugerido testar bases nacionais e incluir variáveis como região, escolaridade e modelo de trabalho. Por fim, destaca-se o uso de Explainable AI para melhorar a transparência e interpretação dos modelos.

BIBLIOGRAFIA

- ARAI, K. et al. **Predição da Rotatividade de Funcionários**. Springer Nature Switzerland AG, 2019.
- BADILLO, S. et al. **An Introduction to Machine Learning**. Clin Pharmacol Ther, 2020.
- MAKHLOUF, K.; ZHIOUA, S.; PALAMIDESSI, C. **Machine learning fairness notions...** Information Processing and Management. 2021.
- VARSHNEY, K. R. **Trustworthy Machine Learning**. 2022.
- WIJAYA, D. **Employee Turnover Dataset**. Kaggle, 2024.
- DALEX. **Fairness in Machine Learning**. Disponível em: <https://dalex.drwhy.ai>