MECAI 2024 III Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Análise Quantitativa da Disparidade Salarial entre Gêneros no Brasil

Luisa Coelho Bolsoni[1], Alexandre Cláudio Botazzo Delbem[2], Kuruvilla Joseph Abraham [3]
ICMC-USP

1 Introdução

A disparidade salarial entre homens e mulheres no Brasil é uma questão ainda presente no mercado de trabalho, mesmo com a implementação de políticas públicas voltadas para a igualdade de gênero. De acordo com o Relatório de Transparência Salarial do Governo Brasileiro (2024), a diferença salarial entre homens e mulheres continua a ultrapassar os 19%. Este trabalho tem como objetivo investigar a contribuição de diferentes fatores para essa desigualdade salarial, utilizando técnicas estatísticas para identificar quais variáveis mais influenciam a probabilidade de estar em uma faixa salarial elevada. O estudo baseou-se nos dados da RAIS, e a Regressão Logística foi escolhida como a técnica principal de modelagem. A escolha dessa abordagem deve-se à sua capacidade de prever variáveis dependentes categóricas, neste caso, gênero e faixa salarial.

2 Metodologia

2.1 Conjunto de Dados

Os dados foram extraídos da RAIS, uma base de dados obrigatória para empregadores no Brasil. A RAIS contém informações detalhadas sobre vínculos empregatícios formais, como escolaridade, idade, setor econômico (CNAE), vínculo e faixa salarial. Para este estudo, focou-se nos dados relacionados às faixas salariais e sexo, complementados por variáveis como escolaridade e idade.

Foi desenvolvida uma função em Python para automatizar a extração anual dos dados, facilitando a atualização e o tratamento dos mesmos. A função foi responsável pela limpeza, remoção de linhas indesejadas, renomeação de colunas e tratamento de valores nulos para que então pudesse ser usada pela regressão logistica.

[1]bolsoni.luisa@usp.br
[2]acbd@icmc.usp.br
[3]abraham@fmrp.usp.br

2.2 Modelos Utilizados

A análise utilizou a Regressão Logística como método principal para prever a faixa salarial e o gênero. Após o ajuste dos modelos, foi utilizada a técnica SHAP (SHapley Additive exPlanations) para avaliar a importância das variáveis explicativas em cada modelo, proporcionando uma interpretação clara do impacto de cada variável.

O SHAP é uma técnica que se baseia nos valores de Shapley da teoria dos jogos, calculando a contribuição marginal de cada variável nas predições do modelo. Ele fornece uma decomposição da predição para cada observação nos dados, mostrando a importância de cada característica.

A equação logística utilizada para modelar a probabilidade de uma variável binária é expressa como:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \tag{1}$$

Onde $P(Y = 1|X)$ representa a probabilidade de estar na faixa salarial mais alta ou de ser homem, dependendo da etapa. As variáveis independentes incluem escolaridade, idade, CNAE e vínculo empregatício.

O SHAP foi aplicado após o treinamento dos modelos, utilizando os comandos da biblioteca SHAP, que permitiram a geração de gráficos de importância e a visualização do impacto das variáveis.

3 Resultados

3.1 Análise exploratória geral dos dados

Na primeira análise exploratória, ao observar a distribuição entre gêneros em cada faixa salarial como é visto na Tabela 1, fica evidente uma inversão: nas faixas salariais mais baixas, há mais mulheres do que homens, enquanto nas faixas mais altas a predominância é masculina. Essa diferença é especialmente visível nas extremidades, como na faixa 1 (de 0 a 0,5 salário) e na faixa 12 (acima de 20 salários), onde a concentração de homens nas faixas superiores se torna bem clara.

Faixa Salarial	**Total**	**Feminino (%)**	**Masculino (%)**
Até 2	8,433,000	4,125,035 (48.9%)	4,307,965 (51.1%)
2 a 7	5,661,135	2,385,579 (42.1%)	3,275,556 (57.9%)
7 a 20	3,220,281	1,397,579 (43.4%)	1,822,702 (56.6%)
Mais de 20	404,710	127,013 (31.4%)	277,697 (68.6%)

Tabela 1: Distribuição de gênero por faixa salarial.

A análise exploratória ao longo do tempo de 2006 a 2022, ilustrada na Figura 1, mostra uma inversão clara nas faixas salariais: até 2 salários mínimos, há maior proporção de mulheres, mas, a partir dessa faixa, os homens predominam.

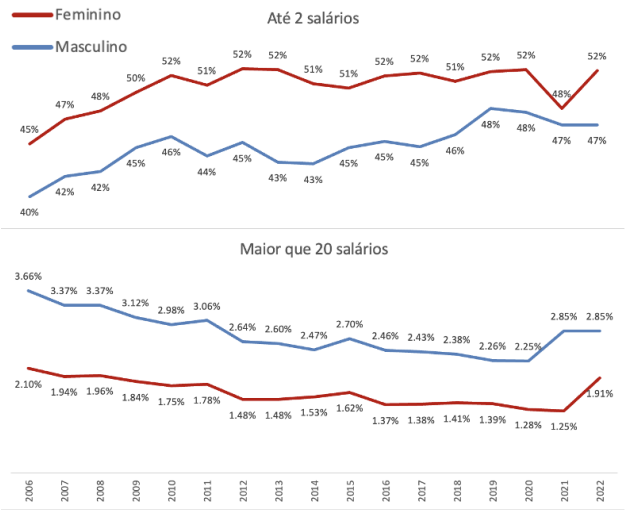

Figura 1: Análise exploratória ao longo do tempo - distribuição de gênero por faixa salarial de 2006 a 2022.

3.2 Importância das Variáveis - Análise SHAP

A análise dos valores SHAP permitiu identificar quais variáveis mais influenciam a predição tanto da faixa salarial quanto do gênero. Em vez de nos concentrarmos nos coeficientes brutos da regressão logística, o SHAP nos oferece uma visualização mais clara e interpretável do impacto de cada variável em cada previsão.

3.3 Previsão da Faixa Salarial

Ao analisar os resultados com a variável dependente sendo a faixa salarial, é possível observar, por meio dos valores de SHAP, que em todas as bases de dados avaliadas o sexo se destaca consistentemente como uma das três variáveis mais importantes para prever o salário. Quando aplicamos o modelo de regressão logística à base de dados do IBGE, que inclui variáveis relacionadas ao setor de atividade econômica, o sexo se consolida como a variável mais relevante na predição salarial. Quando analisados os p-valores resultantes da regressão logística, com a variável dependente sendo a faixa salarial, apenas a variável '30 a 39 anos', dentro da base de idade, não apresentou um p-valor significativo, abaixo de 5%.

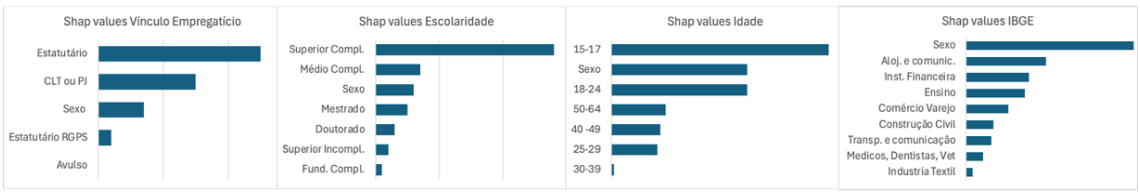

Figura 2: Gráficos SHAP de importância para a predição da faixa salarial.

3.4 Previsão do Gênero

Ao analisar os resultados em que a variável dependente é o sexo, os valores de SHAP indicam que a faixa salarial surge consistentemente como uma das duas variáveis mais relevantes para prever o gênero. Esse resultado revela que a distribuição salarial está fortemente associada ao sexo do indivíduo, sugerindo que há uma clara distinção entre os salários recebidos por homens e mulheres. Quando analisados os p-valores resultantes da regressão logística, com a variável dependente sendo a gênero, todas as variáveis de todas as bases apresentaram p-valores significativos, abaixo de 5%.

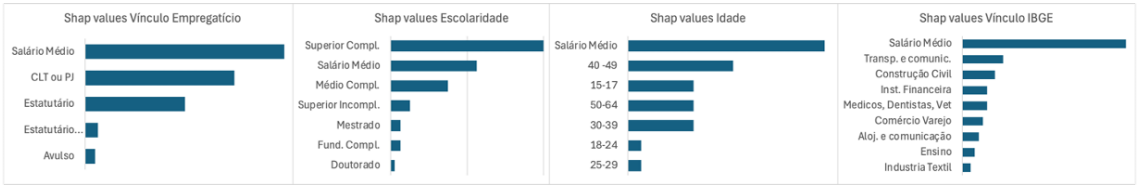

Figura 3: Gráficos SHAP de importância para a predição de gênero.

4 Conclusão

Este estudo demonstra que o gênero é um fator crucial na determinação das faixas salariais no Brasil, com o sexo influenciando fortemente as faixas mais altas. Mesmo considerando outros fatores, como setor econômico e idade.

Referências

[1] Mulheres recebem 19,4% a menos que os homens, aponta 1o Relatório de Transparência Salarial, *Governo Brasileiro*, 2024. Disponível em: https://www.gov.br/secom/pt-br/assuntos/noticias/2024/03/mulheres-ganham-19-4-a-menos-que-os-homens-revela-1o-relatorio-de-transparencia-salarial.

[2] Y. Ansari. Understanding Logistic Regression: A Beginner's Guide, *Medium*, 2024. Disponível em: https://medium.com/@novus$_a fk/understanding - logistic - regression - a - beginners - guide - 73f148866910$.

[3] W. E. Forum. Global Gender Gap Report 2023, *World Economic Forum*, 2023. Disponível em: https://www.weforum.org/reports/global-gender-gap-report-2023. Acesso em: 25 mar. 2024.

[4] A. F. Uceli. Desigualdade Ocupacional no Mercado de Trabalho Formal Brasileiro: Uma Análise da Distribuição de Salários no Século XXI, *Tese (Doutorado)*, Universidade Federal de Minas Gerais, 2023.