Análise Formal dos Resultados

Docente: Willgnner

Discentes: Frederico Lemes Rosa e Maria Clara Ribeiro Di Bragança

Faculdade de Tecnologia Senai de Desenvolvimento Gerencial - FATESG

1. Gráfico de Dispersão e Resíduos

A análise visual dos gráficos de dispersão e de resíduos revela falhas significativas na capacidade preditiva do modelo de Regressão Linear, utilizando a Idade como única variável independente.

1.1 Gráfico de Dispersão: Idade vs. Salário com Linha de Regressão – Evidencia 2

A dispersão vertical ficou alta no eixo Y (Salário) porque os pontos estão muito espaçados para a mesma idade. Por exemplo, entre 40 e 50 anos, há funcionários ganhando R$ 7.000 (desenvolvedores) e outros ganhando mais de R$ 12.000 (gerentes). Isso indica que a idade não é a única, e talvez nem a principal, variável que determina o salário. Fatores como tempo de empresa, nível educacional, dentre outros podem ter um impacto muito maior da variação salarial.

A linha de tendência (linha de regressão vermelha) ficou quase plana, com uma leve inclinação pois o coeficiente R^2 (Coeficiente de determinação) que representa a proporção (ou porcentagem) da variabilidade total da variável dependente (o Salário) que é explicada pelo modelo de Regressão Linear, utilizando a variável independente (a Idade), teve um valor muito baixo, o que significa que para cada ano a mais de idade, o salário aumenta muito pouco. A idade tem uma baixa correlação linear com o salário neste conjunto de dados.

A variação (Banda Rosa) ao redor da linha de regressão representa a margem de erro, ela é bastante larga, o que reforça que a precisão da sua previsão é baixa.

1.2 Gráfico de Resíduos (Erro de Previsão) – Evidencia 2

O Gráfico de Resíduos é a chave para analisar problemas no modelo. Em um modelo de regressão linear ideal, os pontos de erro (resíduos) devem estar

aleatoriamente dispersos em torno da linha horizontal. Entretanto devido ao número insuficiente de dados (Causa Principal) o gráfico mostra apenas alguns pontos isolados. Isso ocorre porque, após a divisão em treino (80%) e teste (20%), a quantidade de dados do conjunto de teste é muito pequena. A consequência de um número muito baixo de pontos no conjunto de teste é que a avaliação do modelo se torna estatisticamente fraca e não confiável. É impossível analisar a aleatoriedade dos erros com tão poucos dados. Já a dispersão irregular dos resíduos, para o pouco que se vê, há um ponto com erro de mais de R$ 2.000(acima da linha) e outro com erro de menos de R$ -2.000(abaixo da linha), o que representa que a magnitude do erro é muito alta em comparação com a média salarial, o que corrobora o alto valor de RMSE (Raiz do Erro Quadrático Médio) nesse caso de 1960.01, e o baixo valor de R^2 que foi de 0.1805 tendo em vista que se fosse próximo de 1 a idade explicaria em grande parte a variação do salário, e quanto mais próximo de 0 (que é o caso) a idade explica muito pouco a variação do salário.

1.3 Conclusão sobre as Causas

Causa	Onde é Vista	Ação Recomendada
1. Quantidade de Dados Insuficiente	Gráfico de Resíduos com poucos pontos.	**Coletar Mais Dados:** Idealmente, você precisaria de centenas de registros para ter uma amostra de teste robusta.
2. Baixa Correlação Linear	Alta dispersão no Gráfico de Dispersão.	**Adicionar Mais Variáveis:** Incluir no modelo fatores como *Experiência, Setor* para tentar explicar a variação salarial.
3. Variância Sendo Explicada por Outros Fatores	A linha de regressão é quase plana.	**Mudar o Modelo:** Se a relação não for linear (por exemplo, o salário aumenta lentamente no início e explode no final da carreira), uma Regressão Polinomial ou outro modelo mais complexo poderia ser considerado.

A principal causa visualmente aparente de que os gráficos estão "estranhos" é o **número insuficiente de dados** para uma análise de teste confiável.

2. Insights

2.1 Insight Principal: A Idade Não é o Fator Determinante

A relação fundamental que aprendemos é que, neste conjunto de dados, a idade tem um impacto marginal sobre o salário, o que confirma uma baixa correlação linear, pois a maior parte dos salários está concentrada em uma faixa estreita (cerca de R$ 12.000). Embora a linha de regressão tenha uma inclinação positiva (sugerindo que a idade aumenta um pouco o salário, a variação vertical (a largura da banda rosa) é muito maior do que o aumento predito pela linha, e o coeficiente R^2 é baixo, indicando que a idade explica uma pequena porcentagem da variação salarial.

2.1 Insight sobre o Erro (Resíduos)

A análise do Gráfico de Resíduos nos dá insights sobre a qualidade da predição e a presença de outliers. Observamos erros (resíduos) de mais de R$ 2.000 tanto acima quanto abaixo da linha Y=0. Isso significa que para algumas idades o modelo superestima o salário (o resíduo negativo, como o ponto próximo a R$ -2.000. Já para outras idades o modelo subestima o salário (o resíduo positivo, como o ponto próximo a R$ 2.000. Ou seja, o modelo é impreciso para prever salários individuais.

2.2 Insight sobre a Estrutura Salarial (Distribuição)

O histograma de salários mostrou uma alta frequência em torno de R$ 12.000 e poucas frequências em outras faixas, isso sugere que a empresa possui uma estrutura salarial altamente padronizada para grande parte dos funcionários (a maioria ocupa o cargo de gerente). E o desvio padrão do salário obtido no df.describe() quantifica essa variação. Como o desvio padrão é baixo em relação a média, isso indica que não há grandes diferenças salariais na empresa, mas devido as conclusões anteriores a idade não é a chave para explicar essas diferenças, sendo outra variável responsável por isso (cargo).

Em resumo, o modelo nos ensinou: A chave para prever salários nessa empresa não está na Idade. Para construir um modelo útil, o foco deve mudar para a coleta e análise de variáveis categóricas como o Cargo, o Setor ou a Experiência, que são os verdadeiros drivers da remuneração.