Data Science Job Salaries 2024

Gabriel Santos de Oliveira Arruda[1], Paulo Henrique Sousa Camargo[2]

FCI – Universidade Presbiteriana Mackenzie

Ciência da Computação - Inteligência Artificial Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

{10388025, 10389453}@mackenzista.com.br

Resumo. *Este artigo apresenta uma análise do dataset "Data Science Job Salaries 2024", que inclui dados sobre ano, nível de experiência, título, salário, moeda, salário em dólar, localização, porcentagem de trabalho remoto e tamanho da empresa. O foco é prever o salário na área da TI nos EUA em empregos de período integral.*

1. Introdução

A área de Data Science tem testemunhado um crescimento significativo nos últimos anos, com a demanda por profissionais qualificados em constante ascensão. Nesse contexto, compreender os padrões salariais e fatores determinantes torna-se crucial para candidatos e empresas. Este estudo se propõe a analisar o conjunto de dados "Data Science Job Salaries 2024", focalizando-se em registros nos Estados Unidos e contratos de tempo integral. A predição do nível de senioridade e localização é o cerne deste projeto, visando fornecer insights valiosos para profissionais e organizações do setor.

1.1. Contextualização

O campo da Ciência de Dados tem se consolidado como uma das áreas mais dinâmicas e em expansão na atualidade. Empresas em diversos setores demandam profissionais qualificados para extrair insights a partir de grandes volumes de dados, o que tem gerado um mercado competitivo e em constante transformação. Neste contexto, entender os fatores que influenciam os salários desses profissionais nos Estados Unidos é crucial para atrair e reter talentos e para orientar decisões estratégicas de carreira e negócios.

1.2. Justificativa

A análise dos padrões salariais e dos fatores determinantes é de extrema relevância para empresas que competem por talentos e para profissionais que buscam se posicionar no mercado. Este estudo se justifica ao oferecer uma visão abrangente sobre a relação entre variáveis, como nível de experiência, localização e trabalho remoto, com a remuneração. Isso permite não apenas identificar tendências, mas também otimizar estratégias de contratação e progressão de carreira.

1.3. Objetivo

Este trabalho tem como objetivo analisar o dataset público "Data Science Job Salaries 2024", que contém informações de contratos de trabalho em tempo integral nos Estados Unidos. A pesquisa foca na criação e avaliação de modelos preditivos para estimar salários anuais em dólares, considerando fatores como nível de experiência,

proporção de trabalho remoto e localização geográfica. Além disso, busca-se compreender as principais variáveis que afetam a remuneração e propor ajustes nos modelos preditivos para melhorar sua eficácia.

1.4. Opção do Projeto

A opção por focar exclusivamente em registros nos Estados Unidos e em contratos de tempo integral foi feita para delimitar o escopo do projeto e garantir a relevância dos resultados para o contexto específico do mercado de trabalho nesse país e para o perfil de profissionais interessados em oportunidades de emprego em tempo integral na área de Data Science.

2. Descrição do problema

O cenário atual do mercado de trabalho em tecnologia, especialmente no campo da ciência de dados, é marcado por uma alta demanda por talentos qualificados. No entanto, tanto candidatos quanto empresas enfrentam desafios ao tentar encontrar o melhor ajuste entre habilidades, expectativas salariais e localização geográfica. Essas questões são complexas, pois estão sujeitas a uma variedade de fatores, como experiência profissional, título do cargo, tamanho da empresa e localização. O conjunto de dados ”Data Science Job Salaries 2024” oferece uma oportunidade única para abordar essas preocupações, fornecendo insights detalhados sobre os padrões salariais e a distribuição geográfica das oportunidades de emprego nos Estados Unidos. Ao realizar análises preditivas sobre os níveis de senioridade e localização, este projeto visa fornecer informações valiosas para profissionais em busca de oportunidades de carreira e empresas procurando atrair e reter talentos no competitivo mercado de tecnologia dos EUA.

3. Dataset

O conjunto de dados utilizado neste projeto foi o “Data Science Job Salaries 2024”, disponível na plataforma Kaggle [999]. Este dataset contém informações detalhadas sobre cargos na área de ciência de dados e tecnologia nos Estados Unidos, exclusivamente para contratos de trabalho em tempo integral. Ele possui 3.000 registros e 10 atributos, descritos a seguir:

- work_year: Ano do registro do trabalho (ex.: 2022, 2023).
- experience_level: Nível de experiência do profissional (Junior, Pleno, Senior, etc.).
- employment_type: Tipo de contrato (tempo integral, meio período, freelancer).
- job_title: Título do cargo ocupado (ex.: Data Scientist, Data Engineer).
- salary: Salário bruto anual informado, na moeda local da empresa.
- salary_in_usd: Salário convertido para dólares, ajustado pela paridade do poder de compra.
- employee_residence: Localização do funcionário (país de residência).
- remote_ratio: Proporção do trabalho realizado remotamente (0%, 50%, 100%).
- company_location: Localização da sede da empresa (país).
- company_size: Tamanho da empresa classificado como pequeno (S), médio (M) ou grande (L).

O atributo salary_in_usd foi selecionado como alvo para a análise preditiva, enquanto os demais atributos foram considerados como potenciais preditores.

Durante a análise exploratória dos dados (EDA), foram realizadas as seguintes observações e ações:

1. Valores Nulos e Duplicados:
    - Foram encontrados registros duplicados e valores nulos em alguns atributos. Para corrigir, os duplicados foram removidos e valores nulos

tratados por exclusão ou preenchimento com a mediana, dependendo da variável.

2. Outliers:
    - Identificaram-se outliers significativos nos salários, especialmente acima de 300.000 dólares anuais. Esses valores foram analisados e, posteriormente, removidos para evitar distorções na modelagem.
3. Distribuição dos Dados:
    - A análise revelou que os salários apresentavam uma distribuição assimétrica e concentrada na faixa entre 50.000 e 150.000 dólares, com menor representatividade em valores extremos.
    - A proporção de trabalho remoto variava significativamente, sendo 100% remoto a modalidade predominante.
4. Correlação entre Atributos:
    - Por meio de um heatmap de correlação, verificou-se que os atributos experience_level e remote_ratio apresentavam uma correlação moderada com o salário. Outros atributos, como company_size, mostraram correlação fraca.

Após o tratamento e pré-processamento, o dataset foi ajustado para incluir apenas os seguintes registros:

- Localização da empresa: Estados Unidos.
- Tipo de contrato: Tempo integral.
- Salários filtrados para valores abaixo de 300.000 dólares.

Por fim, o dataset limpo foi salvo e utilizado para a modelagem preditiva, sendo dividido em conjuntos de treino e teste para validação. As transformações realizadas garantiram que os dados estivessem prontos para aplicação em algoritmos de aprendizado de máquina.

4. Metodologia

A metodologia do projeto foi estruturada em uma sequência de passos bem definidos para alcançar os objetivos propostos. A linguagem de programação utilizada foi Python, na versão 3.9, por sua versatilidade e amplo suporte para bibliotecas de ciência de dados e aprendizado de máquina. Entre as ferramentas empregadas, destacam-se as bibliotecas pandas (versão 1.3.3) para manipulação e análise de dados, numpy (versão 1.21.2) para cálculos matemáticos, seaborn (versão 0.11.2) e matplotlib (versão 3.4.3) para visualização de dados. Para a implementação dos modelos de aprendizado de máquina e avaliação, utilizamos a biblioteca sklearn (versão 0.24.2).

O conjunto de dados utilizado foi "Data Science Job Salaries 2024", disponível no Kaggle, contendo informações relevantes como ano de trabalho (work_year), nível de experiência (experience_level), tipo de emprego (employment_type), salário em dólares (salary_in_usd), proporção de trabalho remoto (remote_ratio), entre outros. Este dataset foi escolhido por sua riqueza de atributos relacionados ao mercado de trabalho em ciência de dados nos Estados Unidos.

O processo começou com o carregamento dos dados a partir de um arquivo CSV e inspecionados, onde foi identificada a presença de valores ausentes, inconsistências e registros duplicados. Esses problemas foram tratados durante a etapa de pré-processamento. Valores nulos foram removidos ou imputados, e registros duplicados foram eliminados para garantir a qualidade dos dados. Outliers, como salários acima de 300.000 dólares, foram excluídos para evitar distorções nas análises.

Após a limpeza, as variáveis categóricas, como o nível de experiência, foram transformadas em valores numéricos utilizando mapeamentos, permitindo sua utilização nos modelos de aprendizado de máquina. Em seguida, foi realizada uma análise exploratória dos dados, com a criação de gráficos como boxplots e heatmaps para visualizar distribuições e correlações entre variáveis. Essa análise revelou que muitas variáveis possuíam correlações fracas com o salário, o que representou um desafio para a construção do modelo preditivo.

Para a modelagem, os dados foram divididos em conjuntos de treino e teste, com proporções de 75% e 25%, respectivamente. Um modelo de regressão linear foi instanciado e treinado no conjunto de treino, utilizando variáveis como work_year, experience_level e remote_ratio para prever o salário. O desempenho do modelo foi avaliado utilizando o coeficiente de determinação (R^2), que indicou que o modelo explicava 13,16% da variância nos salários. Adicionalmente, gráficos de dispersão foram gerados para comparar os valores reais com os previstos, evidenciando as limitações do modelo.

Por fim, os resultados e as visualizações geradas foram salvos para análise e documentação futura. Todo o código, gráficos e documentação foram disponibilizados no repositório público do GitHub, permitindo a replicação do projeto e a contribuição de terceiros para melhorias no modelo. Esta metodologia demonstra o rigor aplicado em todas as etapas, desde a coleta e tratamento dos dados até a análise e avaliação do modelo preditivo.

5. Resultados

5.1. Coeficiente de determinação - R^2

Os resultados obtidos a partir do modelo de regressão linear foram analisados detalhadamente. Após a divisão dos dados em conjuntos de treino e teste (75% para treino e 25% para teste), o modelo foi treinado e as previsões foram realizadas para o conjunto de teste. A precisão do modelo foi avaliada pelo coeficiente de determinação (R^2), que foi calculado em 0.1316. Este valor indica que aproximadamente 13,16% da variância nos salários pode ser explicada pelas variáveis independentes incluídas no modelo. Embora este valor de R^2 seja relativamente baixo, sugerindo uma capacidade limitada do modelo em prever os salários com precisão, ele fornece uma base inicial para melhorias futuras.

```
r2 = r2_score(Y_test, Y_pred)
print("O R2 do modelo é:", r2)
print('Intercepto:', modelo.intercept_)
```

```
O R2 do modelo é: 0.1316022956875531
Intercepto: -6675004.053524757
```

Figura 1. Coeficiente de determinação

5.2. Gráfico de dispersão

O gráfico de dispersão gerado comparando os valores observados com os valores previstos pelo modelo mostra uma dispersão significativa, indicando que o modelo não está prevendo com alta precisão. Os pontos no gráfico representam a relação entre os salários reais e os salários previstos, com a linha preta tracejada indicando onde os valores observados seriam iguais aos valores previstos. A dispersão dos pontos ao longo desta linha sugere que as previsões do modelo tem um nível considerável de erro.

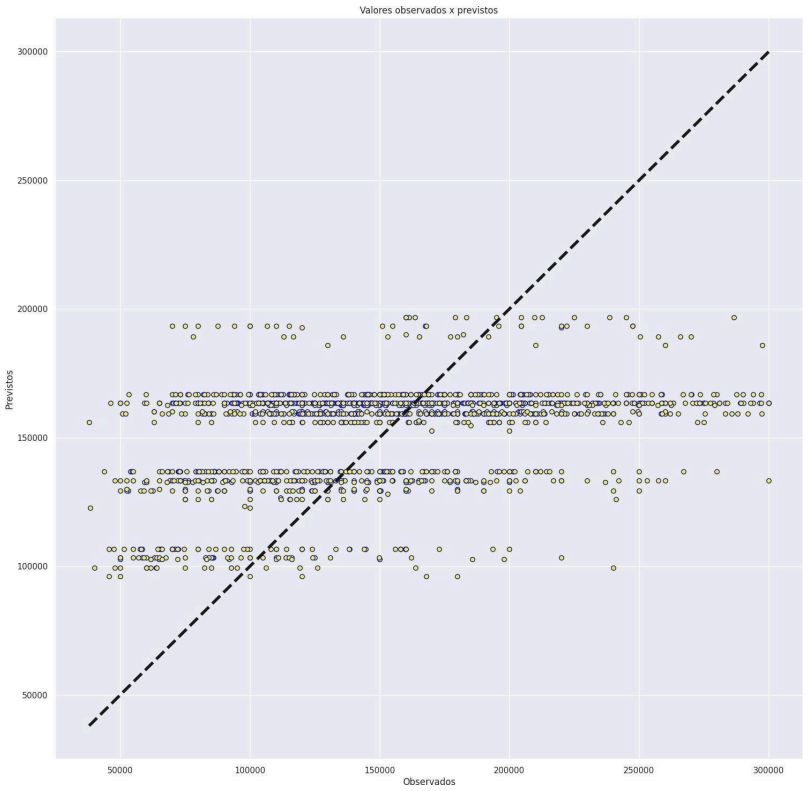

Figura 2. Coeficiente de determinação

5.3. Coeficientes

Além disso, os coeficientes das variáveis independentes no modelo foram analisados. O intercepto do modelo foi de -6675004.05, enquanto os coeficientes das variáveis foram os seguintes: para *work year* foi 3350.67, para *experience level* foi 29982.17, e para *remote ratio* foi -40.37. Esses coeficientes indicam que, para cada unidade de aumento em *work year*, o salário aumenta em média $3350.67, para cada nível de experiência adicional, o salário aumenta em média $29982.17, e para cada unidade de aumento na proporção de trabalho remoto, o salário diminui em média $40.37.

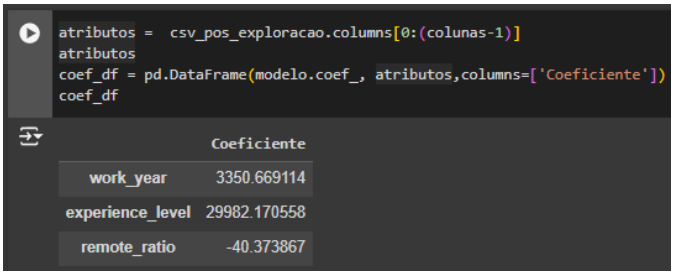

Figura 3. Coeficientes

6. Conclusão

O modelo gerado não se mostrou adequado para uso, uma vez que não aprendeu o suficiente para prever os salários de maneira precisa. Vários fatores podem ter contribuído para esse resultado insatisfatório. Primeiramente, a quantidade limitada de atributos preditores pode ter restringido a capacidade do modelo de capturar as variáveis relevantes. Além disso, a dispersão excessiva dos dados sugere uma grande variabilidade que dificulta a identificação de padrões claros. Por fim, a ausência de correlação significativa entre os atributos utilizados e os salários prejudica a eficácia das previsões, evidenciando a necessidade de uma seleção mais criteriosa dos atributos a serem incluídos no modelo.

7. Link de Acesso ao GitHub

https://github.com/gabriel07099/ProjetoIA

8. Link de Acesso ao Youtube

https://www.youtube.com/watch?v=ttFj_9OZVlA

9. Referências

BILAL, M. Data scientist salary: ultimate 2024 guide. In: BILAL, M., editor. Disponível em: https://www.iu.org/blog/salaries/data-scientist-salary-expectations/. Acesso em: 11 nov. 2024.

REHMAN, F. Data science job salary prediction (glassdoor). In: REHMAN, F., editor. Disponível em: https://www.kaggle.com/code/fahadrehman07/data-science-job-salary-prediction-glassdoor. Acesso em: 11 nov. 2024.

SHAW, A. Data science job salaries 2024. In: SHAW, A., editor. Disponível em: https://www.kaggle.com/datasets/abhinavshaw09/data-science-job-salaries-2024. Acesso em: 11 nov. 2024.

SHAW, A. Data science job salaries EDA. In: SHAW, A., editor. Disponível em: https://365datascience.com/career-advice/data-science-salaries-around-the-world/. Acesso em: 11 nov. 2024.

YOSIFOVA, A. Data science salaries around the world in 2024. In: YOSIFOVA, A., editor. Disponível em: https://365datascience.com/career-advice/data-science-salaries-around-the-world/. Acesso em: 11 nov. 2024.