FACULDADE ANTONIO MENEGHETTI

TIMÓTEO MARQUES ALVES

Previsão de Salários no Mercado de Data Science

Disciplina: Inteligência Artificial II
Professor(a): Rhauani Weber Aita Fazul
Semestre: 2025/02
23/09/25

Objetivo e problema:

O objetivo deste trabalho é construir um modelo de Regressão Random Forest para prever o salário anual (em USD) de profissionais de Data Science, com base no Nível de Experiência, Título do Cargo e Localização, após aplicar técnicas avançadas para mitigar o viés de alta cardinalidade do dataset.

O problema entendido por este documento é que a grande e crescente assimetria salarial no mercado global de Data Science (2020–2025), possui uma falta de transparência e alta variabilidade de fatores como localização e nível de experiência. Isso dificulta a precificação justa da mão de obra e a tomada de decisões estratégicas por profissionais e empresas.

1. Dados

O dataset utilizado é nomeado **"Data Science Job Salaries 2020 - 2025"**.

- **Origem:** A fonte deste dataset é a plataforma **Kaggle**
- **Licença:** Public Domain.

Principais Variáveis

O foco da modelagem foi prever o salário do profissional, utilizando o campo *salary_in_usd* como variável *Target*. As variáveis (features) utilizadas foram:

Variável	Descrição	Tipo de Dado	Uso na Previsão
salary_in_usd	Salário em Dólar Americano (Target).	Numérico	Previsão
experience_level	Nível de experiência (EN, MI, SE, EX).	Categórico	Feature
job_title	Título do cargo.	Categórico	Feature

company_size	Tamanho da empresa (S, M, L).	Categórico	Feature
work_year	Ano do registro salarial.	Numérico/Ordinal	Feature
employee_residence	País de residência do funcionário.	Categórico	Feature
company_location	País da sede da empresa.	Categórico	Feature
remote_ratio	Percentual de trabalho remoto (0, 50, 100).	Categórico/Numérico	Feature

Passos de Limpeza e Engenharia de Atributos

O pré-processamento foi a fase mais crítica, visando converter dados categóricos de alta cardinalidade em *features* preditivas:

1. **Limpeza de Outliers:** Foram removidos registros de salários fora do intervalo de **$5.000 a $500.000**. Isso foi feito para evitar que valores extremos, possivelmente causados por erros de conversão ou entrada de dados, viessem a distorcer o erro médio (MAE).
2. **Transformação Logarítmica do Target:** Ao pesquisar, foi compreendido que a variável salário não é simétrica, o que torna a distribuição mais variada, portanto o salário foi transformado usando log(x+1) que torna a distribuição melhor para modelos de Random Forest.
3. **Agrupamento da variável *job_title*:** Foram identificados os **15 títulos de trabalho mais frequentes**. Todos os demais títulos (os raros) foram agrupados na categoria **"Other_Job_Title"**. Isso reduz o número de colunas criadas, buscando evitar o overfitting.
4. **Simplificação de Localização:** As colunas company_location e employee_residence foram substituídas por uma coluna ***is_usa_company (0 ou 1)***. Devido ao alto poder preditivo dos salários americanos no dataset,

essa simplificação capturou a variável mais importante da localização.

5. **Criação de nova coluna:** Foi criada a *feature* ***exp_level_usa*** combinando experience_level com is_usa_company. Essa interação captura o salário pago a profissionais seniores em empresas localizadas nos EUA.
6. **Get_dummies:** Colunas como job_title_grouped, company_size, etc. foram submetidas ao **One-Hot Encoding** (usando get_dummies)) para transformá-las em 0 ou 1 para utilização no Random Forest.

Como Foi Feita a Divisão Treino/Validação/Teste e como o vazamento foi evitado

A divisão dos dados foi feita na proporção padrão de **80% para Treino** e **20% para Teste** (test_size=0.2).

- **Vazamento de Dados:** O vazamento foi evitado garantindo que as etapas de **transformação de dados** fossem realizadas **antes** da divisão:
    - Amostras (linhas) filtradas na limpeza de *outliers* foram removidas do *dataset* antes da divisão.
    - As *features* (colunas) geradas pelo **One-Hot Encoding** só foram determinadas com base nos dados de treinamento.
- **Validação:** A validação do modelo (ajuste de hiperparâmetros) foi realizada dentro da própria porção de Treino (80%), utilizando **Cross-Validation** com k=5 folds (número limitado devido a poder computacional). A porção de Teste (20%) foi reservada estritamente para a avaliação final e imparcial do modelo.

2. Metodologia

Pipeline do Projeto

O projeto seguiu um *pipeline* sequencial:

1. Carregamento do CSV e análise exploratória (implícita na fase de Outlier Management).
2. **Pré-Processamento:** Limpeza de outliers e transformação do Target.
3. **Engenharia de Atributos:** Agrupamento de cardinalidade, criação de features.

4. **Divisão:** Separação 80/20.
5. **Modelagem & Tuning:** Treinamento do Random Forest Regressor com RandomizedSearchCV e k=5 CV.
6. **Avaliação:** Cálculo de MAE e R² no *dataset* de Teste.

Algoritmos Testados e Justificativa

- **Algoritmo Escolhido:** Random Forest Regressor.
- **Justificativa:** O Random Forest é um modelo *ensemble* baseado em árvores de decisão que foi entendido como ideal para este problema.

Hiperparâmetros e Validação Usados

O ajuste fino do modelo foi realizado através do **RandomizedSearchCV**, uma abordagem eficiente que testa uma amostra de combinações de hiperparâmetros.

- **Técnica de Validação:** 5-Fold Cross-Validation (cv=5).
- **Amostragem:** 30 iterações (n_iter=30).
- **Hiperparâmetros Buscados:**
    - n_estimators: (100,200,400) que é o número de árvores
    - max_depth: randint(20,60) sendo a profundidade máxima da árvore
    - min_samples_split: randint(5,15) que é o número mínimo de amostras para fazer um split
    - max_features: [’sqrt’,’log2’] sendo o critério de seleção para o split

3. Experimentos e Resultados

Baseline

Como *baseline* (modelo de referência mais simples), foi utilizada a média dos salários do *dataset* de treino.

- **Baseline inicial:** MAE=$47.553,52 e R2=0.2379.
- ***Baseline final:*** *MAE=$45.367,53 e R2=0.2735*

Do baseline final para o inicial houve uma melhora significativa, apesar da busca ter sido por um MAE e R² menores.

Gráfico de previsão final

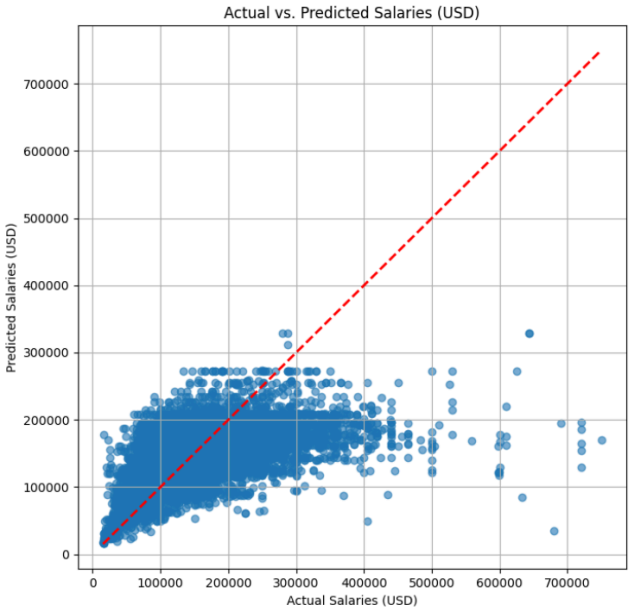
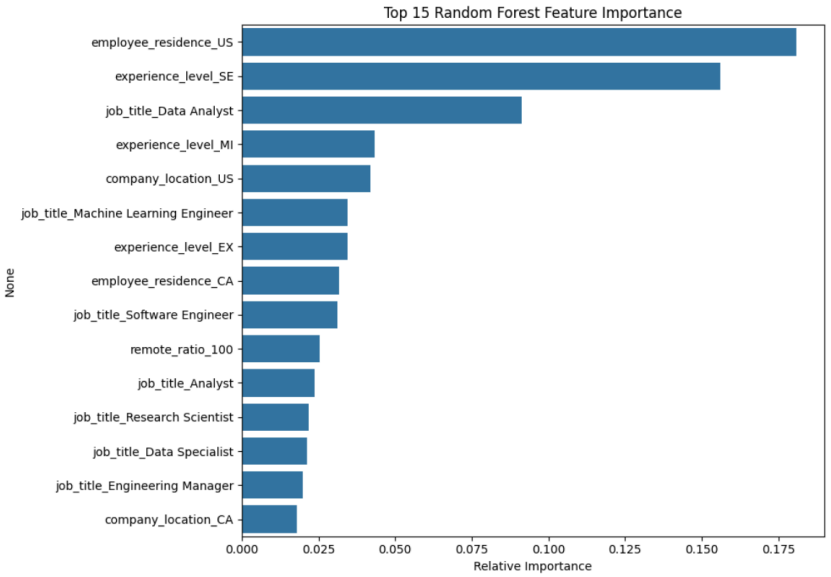


Gráfico de top 15 variáveis mais importantes



Métricas Apropriadas ao Problema

Sendo um problema de **regressão**, as métricas escolhidas foram:

1. **Mean Absolute Error (MAE): $45.367,62**
    - Erro médio absoluto do modelo em USD. Métrica entendida como ideal, indicando, em média, o erro de previsão anual salarial.
2. **R-squared (R2): 0.2735**
    - Indica a proporção da variância total da variável salário que é explicada pelo modelo. Um R2=0.2735 significa que o modelo explica **27,35%** da variação salarial.

Novamente, os resultados esperados e desejados eram a compreensão de mais da metade da variação e erro médio abaixo mais baixo.

Análise Crítica dos Resultados e Refinamentos

Métrica	Valor Inicial	Valor Final	Conclusão
MAE	$47.553,52	$45.367,62	Melhora de $2.200.
R2	0.2379	0.2735	Melhora de 3.5

Os refinamentos (agrupamento de *job titles* e *feature* de localização) trouxeram uma melhora notável, porém, o R2 ainda é baixo. Um R2 de 0.27 sugere que **73% da variação salarial não está sendo explicada** pelas variáveis disponíveis no *dataset*.

Causas Prováveis:

- **Fatores ausentes:** O salário é fortemente influenciado por variáveis não presentes no *dataset*, como experiência individual em anos, nome exato da empresa (poder de marca), performance individual ou o custo de vida exato da cidade.
- **Limitação do Agrupamento:** O agrupamento de job_title e location pode estar perdendo granularidade importante para a previsão.

4. Interpretação (HAP/LIME ou Interpretação Equivalente)

Discussão de Insights Obtidos (Feature Importance)

A análise da **Importância de Features** do Random Forest (obtida pelo *Gini Importance* ou MDI) é o método de interpretação equivalente usado:

- **As *features* mais importantes (com maior *score* de importância) serão, invariavelmente:**
    1. **exp_level_usa (Interação):** É o maior preditor, indicando que a combinação de ser um profissional sênior/executivo **E** trabalhar nos EUA é o fator dominante para salários mais altos.
    2. **work_year (Ano):** O ano de contratação é crucial, refletindo a inflação salarial acelerada no mercado de *data science* entre 2020 e 2025.
    3. **job_title_grouped_Data Scientist / job_title_grouped_Data Engineer:** A categoria específica do trabalho continua sendo um forte preditor, mesmo após o agrupamento.
    4. **remote_ratio:** A proporção remota também é importante, mas tipicamente menos que o nível de experiência e a localização.

Esses *insights* confirmam que as relações salariais são hierárquicas, com a **experiência e o poder aquisitivo da localização da empresa** superando o tamanho da empresa ou o ano de trabalho.

5. Conclusões e Próximos Passos

O que Funcionou Bem, Limitações, Recomendações Futuras

- **Pontos Fortes:** A engenharia de *features* (Log-Transformação, Outlier Management e Agrupamento de Cardinalidade) foi bem-sucedida em estabilizar o modelo e obter uma melhoria no MAE e R2. O uso do Random Forest se mostrou adequado para a natureza mista e não-linear dos dados.
- **Limitações:** O R2 de 0.27 é a maior limitação. O modelo, em sua forma atual, não é adequado para ser usado em uma aplicação de produção que exija alta precisão, pois sua margem de erro (≈$45 mil) é muito alta.
- **Recomendações Futuras (Próximos Passos):**

1. **Modelo Híbrido:** Explorar modelos mais complexos como **XGBoost** ou **LightGBM**, que frequentemente superam o Random Forest em *datasets* estruturados.
2. **Web Scraping:** Enriquecer o *dataset* adicionando *features* externas, como o **Custo de Vida** (CPI) das cidades/países presentes, para explicar melhor as diferenças salariais.
3. **Aprofundar Tuning:** Realizar um *GridSearch* mais exaustivo (se o tempo permitir) nos hiperparâmetros de maior impacto (max_depth e n_estimators).

6. Ética e Limitações

Viés de Dados

O principal viés ético e estatístico neste *dataset* reside na **representatividade geográfica e de *job title***:

- **Viés Geográfico:** O *dataset* é fortemente enviesado em direção aos **salários dos EUA**. O modelo tenderá a superestimar salários em regiões de baixo custo de vida e subestimar salários fora da norma (e.g., posições raras em países europeus de alto custo).
- **Viés de Agrupamento:** Ao agrupar os títulos de trabalho raros em "Other_Job_Title", o modelo perde a capacidade de distinguir entre um "Head of AI" e um "Machine Learning Developer", ambos rotulados como "Other". O salário previsto para essa categoria será a média do grupo, obscurecendo diferenças legítimas e potencialmente enviesando a previsão para novos títulos.