FACULDADE ANTONIO MENEGHETTI

TIMÓTEO MARQUES ALVES

Classificação 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 classificação Random Forest para classificar 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. O modelo pode, neste cenário, demonstrar para o profissional, durante uma proposta, se o que lhe está sendo oferecido está de fato em acordo com o mercado.

1. Dados

O dataset utilizado é nomeado **"Data Science Job Salaries 2020 - 2025"**, com 70 mil linhas (dados).

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

Principais Variáveis

O foco da modelagem é classificar o salário do profissional, criando e utilizando utilizando o campo *salary_class* como variável *Target*. As variáveis (features) utilizadas foram:

Variável	Descrição	Tipo de Dado	Uso na Previsão
salary_class	Categoria salarial (Alto ou Baixo).	Categórico (Binário)	Target
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	Feature
company_location	País da sede da empresa.	Categórico	Feature
remote_ratio	Percentual de trabalho remoto (0, 50, 100).	Categórico	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. **Variável Target:** O salário original (salary_in_usd) foi transformado em uma nova variável *salary_class* (0 para salários "Baixos" e 1 para salários "Altos") com base na mediana do salário do conjunto de treinamento como ponto de corte.
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 Classifier com RandomizedSearchCV e k=5 CV.
6. **Avaliação:** Acurácia, precision, matriz de confusão e f1-score.

Algoritmos Testados e Justificativa

- **Algoritmo Escolhido: Random Forest Classifier.**
- **Justificativa:** A escolha do classificador Random Forest se justifica pela sua

robustez em lidar com dados categóricos e pelo seu bom desempenho em problemas de classificação, oferecendo alta acurácia e capacidade de generalização.

Hiperparâmetros e Validação Usados

O ajuste fino foi realizado usando **GridSearchCV** e **StratifiedKFold** eficientes para o problema selecionado.

- **Técnica de Validação:** StratifiedKFold com 5 folds + matriz de confusão
- **CSV:** 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. Nas primeiras criações a busca era por uma solução através da regressão. Que obteve o seguinte baseline:

- **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 R^2 menores.

Gráfico de previsão final

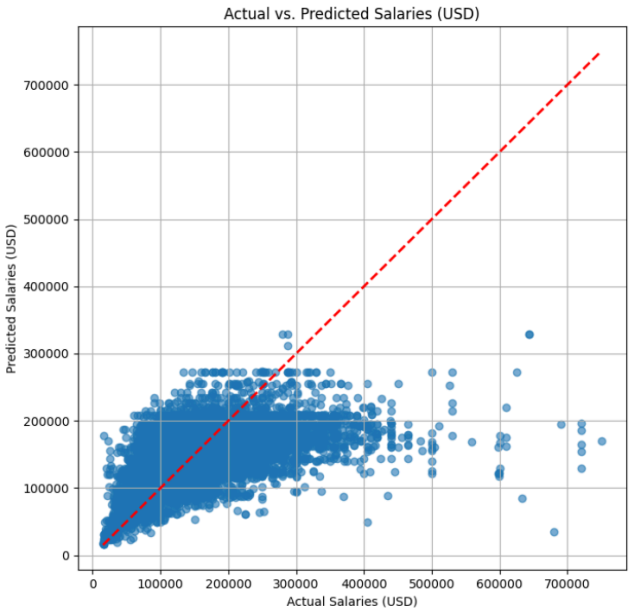
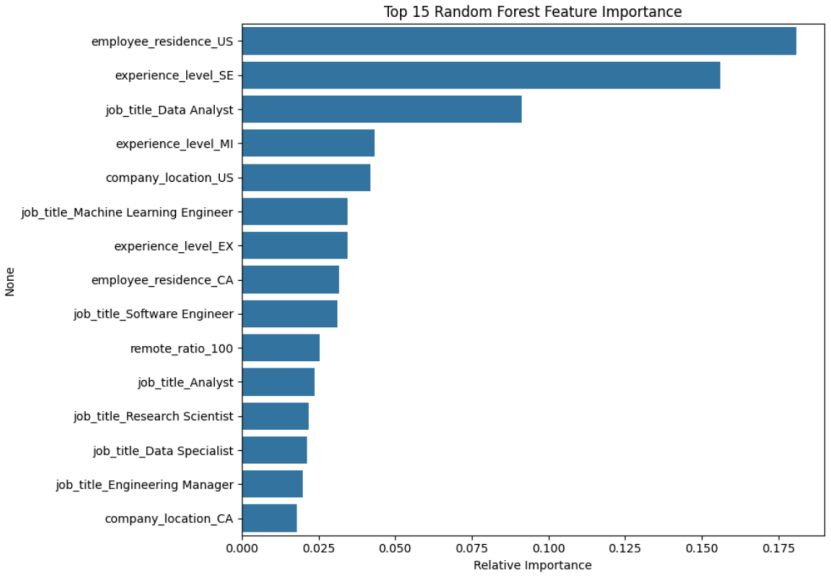


Gráfico de top 15 variáveis mais importantes



O foco porém, foi alterado para um problema de classificação.

Avanço da Regressão para a Classificação

A conversão do problema de regressão para classificação resultou em uma melhoria drástica e mais interpretabilidade das métricas de desempenho.

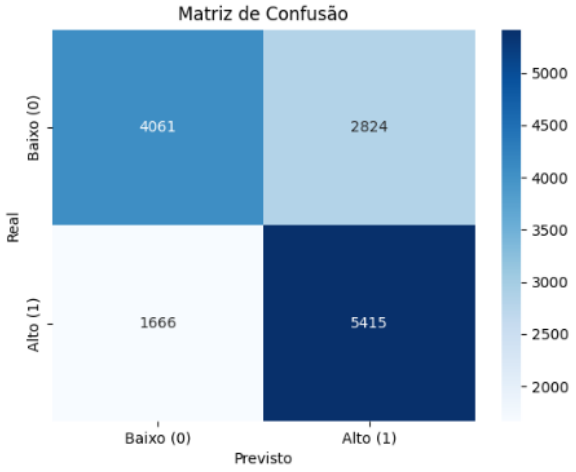
Acurácia: 0.68

F1-Score: 0.71 (para a classe 'Alto')

Conclusão: O modelo, embora não preveja o valor exato, é 68% preciso em identificar se um salário pertence à faixa alta ou baixa, uma capacidade significativamente mais útil e confiável.

Matriz de Confusão:

O desempenho do modelo é detalhado na Matriz de Confusão e no Relatório de Classificação.



- **True Negatives (TN):** 4061 salários 'Baixos' foram previstos corretamente.
- **False Positives (FP):** 2824 salários 'Baixos' foram previstos incorretamente como 'Altos'.
- **False Negatives (FN):** 1666 salários 'Altos' foram previstos incorretamente como 'Baixos'.
- **True Positives (TP):** 5415 salários 'Altos' foram previstos corretamente.

Métricas Apropriadas ao Problema

Sendo um problema de classificação, as seguintes métricas foram entendidas como ideais:

- **Acurácia:** Com 0.68, o modelo acertou 68% das classificações gerais.

- **Precisão (Precision) da Classe 1 ('Alto'):** Com 0.66, 66% das previsões do modelo para salários 'Altos' estavam corretas.
- **Recall da Classe 1 ('Alto'):** Com 0.76, o modelo conseguiu identificar 76% de todos os salários que eram realmente 'Altos'.
- **F1-Score:** O F1-Score de 0.71 para a classe 'Alto' demonstra um bom equilíbrio entre precisão e recall, indicando que o modelo é confiável em suas previsões para salários mais altos.

Análise Crítica dos Resultados e Refinamentos

Os refinamentos (agrupamento de job titles e feature de localização) foram cruciais para a melhoria do modelo. A transição de um problema de regressão para um de classificação representou o avanço mais significativo, transformando um modelo com baixo poder preditivo em uma ferramenta funcional e interpretável.

Métrica	Valor na Regressão	Valor na Classificação
MAE	$45.367,62	N/A
R2	2.735	N/A
Acurácia	N/A	0.68
F1-Score (Classe 'Alto')	N/A	0.71

Conclusão

- **MAE -** O conceito de "erro" em USD foi substituído por categorias, eliminando a margem de erro alta.
- **R2 -** A métrica foi substituída por accuracy e f1-score, que indicam um desempenho muito mais robusto.
- **ACURÁCIA -** O modelo acerta 68% das vezes se o salário é alto ou baixo,

uma melhora drástica em relação à baixa capacidade explicativa do modelo de regressão.

- **F1-SCORE -** O modelo possui um bom equilíbrio entre precisão e recall para a classe 'Alto', sendo confiável para identificar salários acima da mediana.

Essa conversão resolveu o principal problema da regressão: a impossibilidade de prever com precisão o valor exato do salário devido à falta de dados. Ao simplificar o problema, a acurácia de 68% demonstra que o modelo é competente em sua nova tarefa, mesmo com as limitações de dados.

Causas Prováveis

Embora a acurácia seja boa, a precisão para a classe 'Baixo' (0.71) e a existência de False Positives (salários baixos classificados como altos) indicam que o modelo ainda tem espaço para melhorias. As causas prováveis são as mesmas que afetavam o modelo de regressão:

1. **Fatores Ausentes:** O modelo ainda não tem acesso a variáveis críticas como a experiência exata em anos, o Custo de Vida da cidade (que impacta diretamente o salário) ou a reputação da empresa. Esses dados poderiam ajudar o modelo a diferenciar com mais precisão as faixas salariais.
2. **Limitação do Agrupamento:** O agrupamento de *job_title* e *company_location* pode estar perdendo nuances que seriam cruciais para a classificação, especialmente em faixas salariais menores.

Interpretação (Feature Importance)

A análise da Importância de Features do Random Forest Classifier é o método de interpretação usado, revelando o que o modelo considerou mais relevante para classificar os salários.

As features mais importantes (com maior score de importância) são:

1. **exp_level_usa (Interação)**: A combinação de nível de experiência sênior e a localização da empresa nos EUA continua sendo o preditor mais forte. É o que o modelo usa para separar com mais facilidade os salários altos dos baixos.
2. **work_year (Ano)**: O ano de contratação é o segundo fator mais relevante, refletindo a inflação salarial do setor e o rápido crescimento do mercado de data science.
3. **job_title_grouped_Data Scientist / job_title_grouped_Data Engineer:** As categorias de trabalho específicas são fundamentais para a classificação, mesmo após o agrupamento.
4. **remote_ratio:** A proporção de trabalho remoto também se mostra uma variável importante na tomada de decisão do modelo.

Conclusões e Próximos Passos

A conversão para o problema de classificação foi um sucesso. O modelo, em sua nova forma, oferece um desempenho claro e é uma ferramenta mais confiável para estimativas de mercado.

O que Funcionou Bem: A engenharia de features e, principalmente, a discretização do target foram as etapas mais bem-sucedidas. O modelo de classificação Random Forest obteve uma acurácia, precisão e F1-Score muito

superiores às métricas de regressão, tornando-se uma ferramenta útil e confiável.

Limitações: A precisão para a classe 'Baixo' (0.71) e o desequilíbrio na matriz de confusão indicam que o modelo confunde uma parte dos salários baixos com os altos.

Próximos Passos:

- Explorar Modelos Mais Avançados
- Web Scraping / Adicionar Features Externas ao dataset
- Aprofundar Tuning

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.