Course Seven

Google Advanced Data Analytics Capstone

Instructions

Use this PACE strategy document to record your decisions and reflections as a data professional as you work through the capstone project. As a reminder, this document is a resource guide that you can reference in the future and a space to help guide your responses and reflections posed at various points throughout the project.

Portfolio Project Recap

Many of the goals you accomplished in your individual course portfolio projects are incorporated into the Advanced Data Analytics capstone project including:

- Create a project proposal
- Demonstrate understanding of the form and function of Python
- Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
- Demonstrate understanding of how to organize and analyze a dataset to find the “story”
- Create a Jupyter notebook for exploratory data analysis (EDA)
- Create visualization(s) using Tableau
- Use Python to compute descriptive statistics and conduct a hypothesis test
- Build a multiple linear regression model with ANOVA testing
- Evaluate the model
- Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
- Articulate findings in an executive summary for external stakeholders

Project proposal

Análise Preditiva de Retenção de Funcionários na Salifort Motors

Overview

Este projeto visa analisar o conjunto de dados de RH da Salifort Motors para identificar os principais fatores que levam à saída de funcionários. Utilizando técnicas de análise exploratória de dados e a construção de um modelo de machine learning (Random Forest), o objetivo é desenvolver um sistema preditivo que possa prever a probabilidade de um funcionário deixar a empresa. As conclusões e o modelo resultante fornecerão ao departamento de RH insights valiosos e recomendações estratégicas para melhorar as taxas de retenção de talentos.

Milestones	Tasks	PACE stages
Project Proposal	**Definir objetivo (predizer retenção), escopo, métricas de sucesso, identificar stakeholders.**	**Plan**
Exploratory Data Analysis (EDA)	**Carregar dados HR, checar duplicados/nulos, identificar outliers, criar visualizações (boxplots, histograms, scatterplots).**	**Analyze**
Data Preparation & Feature Engineering	**Limpar dataset, criar variáveis derivadas (e.g., "overworked"), codificar categóricas (salary, department).**	**Analyze / Construct**

Model Development	**Construir modelos (Logistic Regression, Decision Tree, Random Forest, XGBoost).**	**Construct**
Model Evaluation	**Comparar modelos com métricas (Accuracy, Precision, Recall, F1, AUC). Escolher o melhor.**	**Construct**
Insights & Recommendations	**Identificar fatores-chave (satisfação, nº de projetos, horas/mês), sugerir políticas de RH.**	**Execute**
Executive Summary & Stakeholder Delivery	**Preparar resumo executivo (1 página), comunicar resultados de forma clara e não técnica, sugerir próximos passos.**	**Execute**

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- **Who is your audience for this project?**

O público-alvo são os executivos de RH e a liderança da Salifort Motors, interessados em compreender os fatores que afetam a retenção de funcionários e em aplicar soluções baseadas em dados para reduzir a rotatividade.

- **What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger business need?**

O objetivo é desenvolver um modelo preditivo capaz de identificar a probabilidade de um funcionário deixar a empresa com base em variáveis como satisfação, desempenho, número de projetos, carga horária e histórico de promoções.

O impacto esperado é fornecer insights acionáveis para reduzir a rotatividade, aumentar a retenção de talentos e, consequentemente, diminuir custos de recrutamento e treinamento, além de melhorar a produtividade geral.

- **What questions need to be asked or answered?**

Quais fatores mais influenciam a saída de funcionários?

Qual o perfil típico de um colaborador que deixa a empresa?

Qual modelo preditivo (regressão ou árvore de decisão/ensemble) apresenta melhor desempenho?

Como a empresa pode atuar preventivamente sobre os fatores identificados?

- **What resources are required to complete this project?**

Dados: HR_capstone_dataset.csv (informações de satisfação, avaliações, projetos, horas, promoções etc.)

Ferramentas: Python (pandas, seaborn, scikit-learn, matplotlib), Jupyter Notebook.

Recursos humanos: Equipe de ciência de dados/analistas para preparar os dados, treinar modelos e interpretar resultados.

Tempo: Janelas de análise e revisões junto aos stakeholders.

- **What are the deliverables that will need to be created over the course of this project?**

Documento PACE Strategy preenchido.

Notebook em Python com todo o fluxo (EDA, modelagem e avaliação).

Resumo executivo (executive summary) em 1 página para stakeholders não técnicos.

Modelo preditivo final (Random Forest/árvore/RegLog), acompanhado de métricas de avaliação.

Visualizações e gráficos explicativos (boxplots, histogramas, feature importance).

Get Started with Python

- **How can you best prepare to understand and organize the provided information?**

Revisar o dicionário de dados do `HR_capstone_dataset.csv`, verificar os tipos de variáveis (numéricas, categóricas, binárias) e garantir que não existam valores ausentes ou duplicados. Organizar os dados em um DataFrame do pandas e criar um plano de limpeza (ex.: tratar outliers, converter variáveis categóricas para dummies).

- **What follow-along and self-review codebooks will help you perform this work?**

Os notebooks exemplares fornecidos pelo curso, além de materiais de prática anteriores sobre EDA, regressão, árvores de decisão e random forest. A documentação oficial do pandas, seaborn e scikit-learn também servirá como referência para análise, visualização e modelagem.

- **What are a couple additional activities a resourceful learner would perform before starting to code?**

Explorar brevemente trabalhos anteriores sobre **retenção de funcionários** para entender quais variáveis costumam ser relevantes.

Criar hipóteses iniciais (ex.: alta carga horária ⟶ maior chance de saída) para guiar a análise.

Definir métricas de avaliação adequadas (acurácia, precisão, recall, F1-score) já no planejamento, garantindo que o modelo seja validado corretamente.

Go Beyond the Numbers: Translate Data into Insights

- **What are the data columns and variables and which ones are most relevant to your deliverable?**

O dataset possui 10 variáveis:

- `satisfaction_level`: nível de satisfação do funcionário (0–1)
- `last_evaluation`: pontuação da última avaliação (0–1)
- `number_project`: número de projetos em que o funcionário participa
- `average_monthly_hours`: média de horas trabalhadas por mês
- `time_spend_company`: anos de empresa
- `work_accident`: ocorrência de acidente no trabalho (0 = não, 1 = sim)
- `left`: alvo, se o funcionário deixou a empresa (0 = ficou, 1 = saiu)
- `promotion_last_5years`: promoção nos últimos 5 anos (0 = não, 1 = sim)
- `department`: departamento de atuação (ex.: vendas, suporte, TI)
- `salary`: nível de salário (low, medium, high)

  Para prever **retenção de funcionários**, as mais relevantes são: `satisfaction_level`, `last_evaluation`, `number_project`, `average_monthly_hours`, `time_spend_company`, `promotion_last_5years`, `salary`.
- **What units are your variables in?**
- `satisfaction_level` e `last_evaluation`: escala contínua de 0 a 1.
- `average_monthly_hours`: horas/mês (inteiro).
- `number_project` e `time_spend_company`: contagem (inteiros).
- `work_accident`, `left`, `promotion_last_5years`: binário (0/1).
- `department` e `salary`: categóricas (texto).
- **What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?**

Presume-se que baixos níveis de satisfação, excesso de projetos e longas jornadas mensais aumentam a probabilidade de saída. Além disso, salários baixos podem estar associados a maior rotatividade.

- **Is there any missing or incomplete data?**

A inspeção inicial (`.isna().sum()`) mostra que não há valores ausentes.

- **Are all pieces of this dataset in the same format?**

Sim, mas variáveis categóricas como `department` e `salary` precisam ser convertidas em dummies/encoding para uso em modelos de ML.

- **Which EDA practices will be required to begin this project?**

Estatísticas descritivas (média, mediana, dispersão).

Visualizações (boxplots, histogramas, scatterplots).

Identificação de outliers em `average_monthly_hours` e `time_spend_company`.

Análise de correlação entre variáveis preditoras e o alvo `left`.

The Power of Statistics

- **What is the main purpose of this project?**

O principal objetivo é prever a probabilidade de um funcionário deixar a empresa e identificar os fatores mais relevantes que afetam a retenção. Com isso, a Salifort Motors poderá adotar políticas mais eficazes de RH e reduzir a rotatividade.

- **What is your research question for this project?**

Quais características dos funcionários (ex.: nível de satisfação, carga horária, número de projetos, salário) estão mais associadas ao desligamento voluntário ou involuntário?
E até que ponto conseguimos prever corretamente a saída de um funcionário com base nessas variáveis?

- **What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn't use random sampling?**

A amostragem aleatória garante que o modelo represente de forma justa toda a população de funcionários, sem privilegiar um grupo específico.
Se não usarmos random sampling, podemos introduzir sampling bias. Por exemplo:

1. Se o conjunto de treino tiver desproporcionalmente muitos funcionários de um único departamento (ex.: vendas), o modelo pode aprender padrões enviesados e não generalizar bem para outros setores (ex.: TI, RH).

2. Outro viés seria treinar o modelo apenas com funcionários de salários baixos, levando a previsões incorretas para funcionários de salários médios ou altos.

Regression Analysis: Simplify Complex Data Relationships

- **Who are your stakeholders for this project?**

Os stakeholders são a liderança e os gestores do departamento de Recursos Humanos (RH) da Salifort Motors.

- **What are you trying to solve or accomplish?**

O objetivo é identificar os principais fatores que influenciam a rotatividade de funcionários e construir um modelo preditivo para antecipar quais funcionários têm maior probabilidade de deixar a empresa.

- **What are your initial observations when you explore the data?**

Observa-se que há um número significativo de funcionários que deixaram a empresa. As colunas contêm uma mistura de dados numéricos (como nível de satisfação) e categóricos (como departamento e salário), e existem muitas linhas duplicadas que precisam ser tratadas.

- **What resources do you find yourself using as you complete this stage? (Make sure to include the links.)**

Os principais recursos são as documentações oficiais das bibliotecas Python utilizadas:

Pandas: https://pandas.pydata.org/docs/

Matplotlib: https://matplotlib.org/stable/contents.html

Seaborn: https://seaborn.pydata.org/

Scikit-learn: https://scikit-learn.org/stable/

- **Do you have any ethical considerations in this stage?**

Sim. É fundamental garantir o anonimato dos dados dos funcionários para proteger sua privacidade. As conclusões não devem ser usadas para punir, mas sim para criar políticas de retenção mais justas e eficazes, evitando a criação de vieses contra determinados grupos ou departamentos.

The Nuts and Bolts of Machine Learning

- **What am I trying to solve?**

Estou tentando resolver um problema de classificação binária: prever se um funcionário vai sair da empresa (`left = 1`) ou se vai ficar (`left = 0`).

- **What resources do you find yourself using as you complete this stage?**

Além das documentações já mencionadas, a documentação do `GridSearchCV` da Scikit-learn foi essencial para a otimização de hiperparâmetros: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

- **Is my data reliable?**

Os dados parecem ser confiáveis após a limpeza, especialmente após a remoção de mais de 3.000 entradas duplicadas. A remoção dessas duplicatas aumenta a integridade do dataset para a modelagem.

- **Do you have any additional ethical considerations in this stage?**
- Sim. Ao construir o modelo, é crucial verificar se ele não está discriminando injustamente com base em variáveis como o departamento. O objetivo é entender o comportamento e não criar um sistema que reforce estereótipos existentes.
- **What data do I need/would I like to see in a perfect world to answer this question?**

Em um mundo ideal, seria útil ter mais informações, como dados demográficos (idade, gênero), avaliações de desempenho qualitativas (feedback de gestores), informações sobre deslocamento diário e histórico de salários e promoções.

- **What data do I have/can I get?**

Temos acesso ao dataset `HR_capstone_dataset.csv`, que contém informações sobre satisfação, avaliação, carga de trabalho, tempo de casa, acidentes, promoções, departamento e faixa salarial.

- **What metric should I use to evaluate success of my business objective? Why?** As métricas chave são **Recall** e **F1-Score**.

A Acurácia por si só pode ser enganosa se o dataset for desbalanceado. O **Recall** é especialmente importante porque o custo de não identificar um funcionário que vai sair (falso negativo) é muito maior do que o custo de intervir com um funcionário que não ia sair (falso positivo).

Data Project Questions & Considerations

PACE: Analyze Stage

Get Started with Python

- **Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?**

Sim. A análise exploratória inicial mostrou que variáveis como `satisfaction_level`, `time_spend_company` e `number_project` têm uma forte correlação com a saída de funcionários, indicando que os dados são suficientes para construir um modelo preditivo robusto.

Go Beyond the Numbers: Translate Data into Insights

- **What steps need to be taken to perform EDA in the most effective way to achieve the project goal?**

Os passos foram: limpeza de dados (remoção de duplicatas), análise univariada para entender cada variável, análise bivariada para comparar cada variável com a variável alvo (`left`) usando visualizações, e uma análise de correlação para entender a relação entre as variáveis numéricas.

- **Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?**

Não foi necessário adicionar mais dados. A principal estruturação foi a remoção de linhas duplicadas. Nenhuma outra filtragem foi necessária para esta análise.

- **What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?**

Para o RH, gráficos claros e diretos são os melhores. Gráficos de barras (countplots) para comparar categorias, mapas de calor (heatmaps) para correlações e scatterplots para visualizar a relação entre satisfação e avaliação são muito eficazes.

The Power of Statistics

- **Why are descriptive statistics useful?**

Estatísticas descritivas (como média, mediana, desvio padrão) são úteis porque resumem as principais características de um conjunto de dados. Elas nos dão uma visão rápida e quantitativa da distribuição dos dados, ajudando a identificar padrões, a normalidade e possíveis outliers antes de realizar análises mais complexas.

- **What is the difference between the null hypothesis and the alternative hypothesis?**

A hipótese nula (H0) é uma afirmação padrão de que não existe efeito ou relação entre as variáveis que estão sendo estudadas. É a hipótese que tentamos refutar. A hipótese alternativa (H1 ou Ha) é a afirmação oposta, sugerindo que existe, sim, um efeito ou uma relação. É a conclusão que aceitamos se conseguirmos rejeitar a hipótese nula com evidências estatísticas suficientes.

Regression Analysis: Simplify Complex Data Relationships

- **What are some purposes of EDA before constructing a multiple linear regression model?**

Antes de construir um modelo de regressão, a Análise Exploratória de Dados (EDA) é crucial para:

1. Verificar a linearidade entre as variáveis independentes e a variável dependente.
2. Identificar outliers que podem influenciar indevidamente o modelo.
3. Verificar a multicolinearidade (alta correlação entre variáveis independentes).
4. Avaliar se os resíduos do modelo seguem uma distribuição normal, uma premissa importante da regressão linear.

- **Do you have any ethical considerations in this stage?**

Sim. Na fase de análise, é importante estar ciente de possíveis vieses nos dados. Devemos garantir que as conclusões tiradas não reforcem estereótipos negativos sobre grupos de funcionários (por exemplo, associar um departamento a um baixo desempenho sem uma causa clara). A privacidade dos dados dos funcionários deve ser sempre a principal prioridade.

The Nuts and Bolts of Machine Learning

- **What am I trying to solve? Does it still work? Does the plan need revising?**

O objetivo continua o mesmo: construir um modelo de classificação para prever a rotatividade de funcionários. O plano se mostrou eficaz durante a EDA, confirmando que os dados continham sinais preditivos fortes, portanto, não houve necessidade de revisão.

- **Does the data break the assumptions of the model? Is that ok, or unacceptable?**

O modelo escolhido, Random Forest, é uma grande vantagem aqui. Por ser um algoritmo baseado em árvores de decisão, ele não exige premissas rígidas sobre a distribuição dos dados, como linearidade ou normalidade. Portanto, o formato dos nossos dados é perfeitamente aceitável para este tipo de modelo.

- **Why did you select the X variables you did?**

  Selecionei todas as variáveis disponíveis como preditoras (X) porque o Random Forest é robusto e consegue lidar com um grande número de features, identificando internamente quais são as mais importantes. Isso nos permite extrair o máximo de informação dos dados sem ter que selecionar manualmente as variáveis de antemão.

- **What are some purposes of EDA before constructing a model?**

  A EDA antes da modelagem ajuda a entender os padrões dos dados, formular hipóteses sobre quais variáveis serão importantes, informar as etapas de pré-processamento (como a necessidade de transformar variáveis categóricas) e detectar problemas como dados duplicados ou outliers.

- **What has the EDA told you?**

  A EDA nos mostrou claramente que o nível de satisfação, o tempo de empresa e o número de projetos são os fatores mais fortemente associados à decisão de um funcionário deixar a empresa. Ela validou que o problema é "solucionável" com os dados que temos.

- **What resources do you find yourself using as you complete this stage?**

  Os recursos mais utilizados foram as documentações online das bibliotecas Python, especialmente Scikit-learn para a modelagem, e Pandas e Seaborn para a manipulação e visualização de dados.

- **Do you have any ethical considerations in this stage?**

  Sim. A principal consideração ética é garantir que o modelo não seja "caixa-preta". Precisamos ser capazes de interpretar *por que* o modelo está fazendo certas previsões (usando técnicas como a importância de features) para garantir que ele seja justo e não esteja baseando suas decisões em vieses indesejados.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

Do any data variables averages look unusual?

Não, as médias das variáveis de dados parecem estar dentro de um intervalo esperado. Por exemplo, o `satisfaction_level` (nível de satisfação) tem uma média de 0.61, o que é razoável em uma escala de 0 a 1. Não há valores que sugiram erros de entrada ou anomalias graves.

How many vendors, organizations or groupings are included in this total data?

Os dados pertencem a uma única organização, a Salifort Motors. Dentro dela, os funcionários estão agrupados em 10 departamentos diferentes (como `sales`, `technical`, `support`, etc.).

Go Beyond the Numbers: Translate Data into Insights

- **What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?**

Para atingir os objetivos, foi necessário construir um algoritmo de machine learning de classificação (Random Forest Classifier). As saídas essenciais foram um relatório de classificação (com métricas como precisão e recall), uma matriz de confusão para visualizar o desempenho e um gráfico de barras mostrando a importância de cada feature.

- **What processes need to be performed in order to build the necessary data visualizations?**

Os processos incluíram a limpeza e preparação dos dados com a biblioteca Pandas, seguida pelo uso das bibliotecas Matplotlib e Seaborn para criar os gráficos, como `countplot`, `heatmap` e `barplot`.

- **Which variables are most applicable for the visualizations in this data project?**

As variáveis mais importantes para visualização foram `satisfaction_level`, `time_spend_company`, `number_project` e a variável alvo, `left`, pois a Análise Exploratória de Dados (EDA) mostrou que elas tinham a relação mais forte com a rotatividade de funcionários.

- **Going back to the Plan stage, how do you plan to deal with the missing data (if any)?**

O plano era verificar a existência de dados ausentes. Neste projeto específico, o dataset não continha valores ausentes, então nenhuma ação de imputação foi necessária. O principal problema de limpeza foram os dados duplicados, que foram removidos.

The Power of Statistics

- **How did you formulate your null hypothesis and alternative hypothesis?**

Embora não tenhamos realizado um teste de hipótese formal neste projeto, um exemplo seria:

- **Hipótese Nula (H0):** Não há diferença na média do nível de satisfação entre os funcionários que saíram e os que ficaram.
- **Hipótese Alternativa (Ha):** A média do nível de satisfação dos funcionários que saíram é significativamente diferente da dos que ficaram.

- **What conclusion can be drawn from the hypothesis test?**

Com base na nossa EDA, a conclusão seria rejeitar a hipótese nula. A visualização dos dados mostrou uma clara diferença no nível de satisfação entre os dois grupos, indicando que a satisfação é um fator estatisticamente significativo na rotatividade.

Regression Analysis: Simplify Complex Data Relationship

- **Do you notice anything odd?**

Não foi notado nada particularmente estranho que pudesse comprometer o modelo. Os padrões identificados na EDA, como a maior taxa de saída para funcionários com 3-5 anos de empresa, são consistentes com cenários de RH do mundo real.

- **Can you improve it? Is there anything you would change about the model?**

O modelo atual (Random Forest) já apresenta um desempenho excelente, com acurácia acima de 98%. Uma possível melhoria seria testar algoritmos de boosting, como XGBoost ou LightGBM, que às vezes podem superar o Random Forest. Outra abordagem seria a engenharia de novas features, combinando variáveis existentes.

The Nuts and Bolts of Machine Learning

- **Is there a problem? Can it be fixed? If so, how?**

Não houve um problema significativo com o modelo. Ele treinou bem e generalizou de forma eficaz os dados de teste, indicando que não há um problema grave de overfitting.

- **Which independent variables did you choose for the model, and why?**

Foram escolhidas todas as variáveis independentes disponíveis no dataset. A razão é que o Random Forest é eficaz em lidar com um grande número de variáveis e pode determinar internamente quais são as mais preditivas, evitando a necessidade de uma seleção manual complexa.

- **How well does your model fit the data? (What is my model's validation score?)**

O modelo se ajusta muito bem aos dados. A pontuação de validação (acurácia no conjunto de teste) foi de aproximadamente 98.6%, e o F1-score para a classe "saiu" foi de 96%, o que indica um modelo altamente preciso e robusto.

- **Can you improve it? Is there anything you would change about the model?**

Como mencionado anteriormente, embora o modelo seja muito bom, testes com outros algoritmos (XGBoost) ou a criação de novas variáveis (engenharia de features) poderiam levar a pequenas melhorias incrementais no desempenho.

- **Do you have any ethical considerations in this stage?**

Sim. É fundamental garantir que o modelo seja interpretável e justo. Devemos evitar que ele seja uma "caixa-preta" e garantir que suas previsões não sejam usadas para discriminar ou tomar ações punitivas contra funcionários, mas sim para criar um ambiente de trabalho melhor.

Data Project Questions & Considerations

PACE: Execute Stage

Get Started with Python

- **Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing an exploratory data analysis?** Eu recomendaria investigar a razão da grande quantidade de dados duplicados (mais de 3.000). Isso poderia indicar um problema no processo de coleta de dados que precisa ser corrigido para garantir a qualidade de análises futuras.
- **What data initially presents as containing anomalies?** A anomalia mais evidente não estava nos valores em si, mas na estrutura dos dados: a presença de um grande número de linhas completamente duplicadas.
- **What additional types of data could strengthen this dataset?** Dados qualitativos de entrevistas de desligamento, avaliações de desempenho mais detalhadas (não apenas uma nota), informações sobre o tempo de deslocamento até o trabalho e dados demográficos poderiam enriquecer a análise e fortalecer o modelo.

Go Beyond the Numbers: Translate Data into Insights

- **What key insights emerged from your EDA and visualizations(s)?** Os principais insights foram: 1) O baixo nível de satisfação é o principal indicador de que um funcionário vai sair. 2) O período de 3 a 5 anos na empresa é um ponto crítico de retenção. 3) Cargas de trabalho muito altas ou muito baixas (medidas pelo número de projetos) aumentam a chance de saída.
- **What business recommendations do you propose based on the visualization(s) built?** Recomendo implementar pesquisas de satisfação mais frequentes, criar um plano de carreira claro para funcionários com mais de 3 anos de casa e revisar a distribuição de projetos para evitar o esgotamento (burnout) ou o tédio.
- **Given what you know about the data and the visualizations you were using, what other questions could you research for the team?** Poderíamos pesquisar: "Funcionários com salários mais altos têm um nível de satisfação maior?" ou "Existe algum departamento com uma taxa de rotatividade desproporcionalmente alta em comparação com os outros?".
- **How might you share these visualizations with different audiences?** Para uma audiência executiva, eu usaria gráficos simples e de alto impacto em uma apresentação de slides. Para uma equipe de análise de dados, eu compartilharia o Jupyter Notebook completo, permitindo uma exploração mais profunda e técnica.

The Power of Statistics

- **What key business insight(s) emerged from your A/B test?** Este projeto não incluiu um A/B test. Ele se concentrou na análise de dados observacionais e na construção de um modelo preditivo.
- **What business recommendations do you propose based on your results?** As recomendações são: monitorar ativamente a satisfação dos funcionários, investir em planos de retenção para funcionários de médio prazo e gerenciar a carga de trabalho de forma mais eficaz.

Regression Analysis: Simplify Complex Data Relationships

- **To interpret model results, why is it important to interpret the beta coefficients?** Em um modelo de regressão, os coeficientes beta indicam a magnitude e a direção da relação entre cada variável independente e a variável dependente. Interpretá-los é crucial para entender *quão forte* é o impacto de cada fator e se esse impacto é positivo ou negativo.
- **What potential recommendations would you make to your manager/company?** As recomendações são as mesmas derivadas do modelo de machine learning: focar em aumentar a satisfação, desenvolver planos de carreira e equilibrar a carga de trabalho.
- **Do you think your model could be improved? Why or why not? How?** Sim, todo modelo pode ser melhorado. Embora nosso modelo de Random Forest seja muito preciso, ele poderia ser aprimorado com mais dados (como os mencionados anteriormente) ou testando algoritmos mais complexos como o XGBoost, que poderiam capturar padrões ainda mais sutis.

The Nuts and Bolts of Machine Learning

- **What key insights emerged from your model(s)?** O principal insight do modelo foi a confirmação quantitativa de que o `satisfaction_level` é o fator mais importante, seguido por `time_spend_company` e `number_project`. Isso dá ao RH um foco claro sobre onde direcionar seus esforços de retenção.
- **What are the criteria for model selection?** Os critérios foram a performance preditiva (medida pelo F1-Score e Recall) e a capacidade de interpretação (a importância das features).
- **Does my model make sense? Are my final results acceptable?** Sim, o modelo faz muito sentido, pois suas conclusões estão alinhadas com as práticas de RH. Os resultados são mais do que aceitáveis para implementação, dada a alta precisão alcançada.
- **Were there any features that were not important at all? What if you take them out?** A variável `acidente_trabalho` teve a menor importância. Removê-la provavelmente não diminuiria significativamente o desempenho do modelo e poderia torná-lo um pouco mais simples.
- **Given what you know about the data and the models you were using, what other questions could you address for the team?** Poderíamos abordar: "Podemos prever o *nível de satisfação* de um funcionário com base em outras variáveis?" ou "Podemos criar clusters de funcionários com diferentes perfis de risco de saída?".

- **What resources do you find yourself using as you complete this stage?** Os recursos mais úteis foram a documentação da biblioteca Scikit-learn para entender os parâmetros do modelo e as métricas de avaliação.
- **Is my model ethical?** O modelo em si é uma ferramenta matemática e, portanto, neutro. Sua aplicação se torna uma questão ética. Se usado para melhorar proativamente o ambiente de trabalho e apoiar os funcionários, é ético. Se usado para punir ou discriminar, não é.
- **When my model makes a mistake, what is happening? How does that translate to my use case?** O erro mais crítico é o **falso negativo**, que ocorre quando o modelo prevê que um funcionário vai ficar, mas na realidade ele sai. Isso se traduz em uma oportunidade perdida de intervir e reter um talento. O nosso modelo foi otimizado para minimizar esse tipo de erro, conforme mostrado pelo seu alto recall.