1. **Insights retirados das entrevistas**

**1.1) Regularização**

É uma técnica para evitar o overfitting, ou seja, evitar que o modelo se adapte demais aos dados de treino, de modo a apenas decorar os dados que estão nele. A regularização adiciona um termo de penalidade na função custo(Loss function), evitando que o modelo tente minimizar a Loss Function depois de passar de um ponto “Saudável”, corrigindo o aprendizado

**1.2) Modelos Ensemble**

É a técnica de combinar abordagens para obter um melhor resultado, o Random Forest, por exemplo, combina várias árvores simples para ter um modelo mais preciso. O ensemble pode ser bagging e boosting, o primeiro está relacionado a fazer várias previsões, como no RF, e a partir da média dessas previsões, setar o valor real. O segundo é usado quando pegamos a previsão de um modelo e colocamos como input de outro, fazendo com que cada modelo aprenda com seu antepassado.

**1.3) Como a primeira pergunta de uma árvore de decisão é determinada?**

Uma árvore de decisão pode ser resumida em um fluxograma de perguntas que levam a uma resposta final, sendo assim, temos que determinar a ordem das perguntas. A primeira é aquela que melhor separa uma base de dados, por exemplo, queremos separar bons e maus pagadores, vamos sempre escolher a pergunta que deixe um lado o mais próximo possível de 100% dos bons pagadores e o outro com 100% dos maus pagadores. Na prática, determinamos isso através do índice de Gini ou Entropia

1. **Regressão Linear**

**2.1) Derivando os parâmetros**

Como nosso objetivo é minimizar a Loss Function para encontrar os melhores parâmetros, podemos usar a derivada da função para encontrar a função que erra menos.

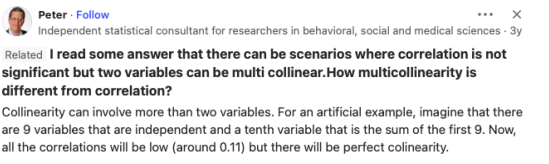
**2.2) Premissas, Inferências e Predição**

Uma *premissa* é uma suposição fundamental ou condição que é feita antes de uma análise ou modelagem de dados, ajudando a definir o contexto e as restrições de um problema, violar essas premissas podem levar a resultados imprecisos. Algumas premissas importantes são: Linearidade nos parâmetros: as relações entre as variáveis independentes(features) e a target é linear, observações independentes entre si: Assume que a parte dos dados usadas para treinar e pra testar são independentes entre si, variância constante nos erros: assume que a variância dos erros do modelo é constante em todas as faixas de valores, normalidade dos resíduos: A diferença entre as previsões e valores reais seguem uma distribuição normal, multicolinearidade não-perfeita: Evitar a multicolinearidade não perfeita ou moderada é importante para que não tenham variáveis preditoras muito relacionadas entre si, pois isso afetaria a forma na qual o modelo calcula o quanto uma variável é importante na hora de inferir o resultado. Apesar disso, é importante ter cautela na hora de falar que o modelo não vai funcionar caso não atenda essas premissas, uma vez que isso é muito relativo, não vai funcionar no sentido de Loss Function? previsão sem sentido? tudo? Como tudo em machine learning, DEPENDE!

**2.3) Correlação X Multicolinearidade**

A correlação é uma medida de força de direção entre duas variáveis independentes, enquanto a multicolinearidade ocorre quando duas ou mais variáveis independentes em um modelo estão altamente correlacionadas entre si. Em diversas fontes, é dito que uma alta correlação implica em colineraridade, ou o contrário, no caso do autor, ele discorda. Um exemplo é que quando temos 9 colunas prevendo uma outra coluna, que seria a soma das 9 anteriores, por exemplo, todas as correlações iam ser muito baixas(por volta de 0.11) porém ia ter uma colinearidade perfeita

1. **Data Science na prática**

Nessa seção, o autor trabalha em cima de uma base de dados sobre Chum de bancos(quando o cliente sai da plataforma) e as primeiras perguntas que ele faz são referentes ao mapeamento do cliente: o que faz o cliente sair da plataforma, idade? gênero? Ou seja, queremos mapear quais clientes saem. Nessa parte, não tem um passo a passo muito definido, o ideal é ir matando a curiosidade e a partir disso extrair insights importantes.

Quando for a hora de fazer a análise exploratória de dados é interessante separar a análise entre categóricas e numéricas

**Algumas escolhas de gráficos tradicionais:**

* *Distribuição de variáveis contínuas:* Boxplot, Histograma e swarmplot
* *Distribuição de variáveis categóricas:* Gráfico de barras com a frequência de cada categoria
* *Relação entre duas variáveis contínuas:* Gráfico de dispersão
* *Relação entre uma variável categórica e uma contínua:* Utilizar boxplot para as diferentes categorias, fazendo um subplot para cada possibilidade de valor que a categoria pode assumir
* *Relação entre duas variáveis categóricas:* Uma tabela cruzada com a contagem