7.3 - Introdução a regressão linear (parte 2)

00:00:00:00 - 00:00:01:43
Vamos plotar um gráfico dos valores reais versus os valores preditos pelo modelo. Vamos importar a biblioteca matplotlib, né? Import matplotlib ponto pyplot as plt. E vamos plotar um gráfico começando criando uma nova figura. Uma figura... Vamos colocar um fig size também para ter um tamanho legal e tal. plt ponto figure e aqui dentro a gente coloca um fig size. Vamos colocar aqui de 10 polegadas por 6, e depois se ficar muito ruim a gente modifica esses números. Vamos plotar um gráfico de dispersão dos valores reais contra os valores previstos, e definimos a transparência dos pontos por 0.5, que é aquele alfa. Vamos lá, plt ponto scatter, que são os gráficos de dispersão. O primeiro parâmetro é o ytest, o segundo parâmetro é o ypred, E o terceiro parâmetro é aquele alfa que é a transparência dos pontinhos. Ok. Bom, e aí a gente vai colocar os rótulos para ficar bacaninha. X label é o valor real e o Y label é o valor predito. E vamos colocar também um título, plt ponto title. Vamos colocar dispersão dos dados. Depois a gente pensa num nome melhor, hein.

00:00:01:43 - 00:00:03:22
Bom, e aí a gente vai colocar um plt ponto plot que vai definir alguns pontos de início ao fim da linha para poder criar uma reta bonitinha no nosso gráfico. Vai ser plt ponto plot, eu vou colocar aqui a função e depois eu vou discutir no ponto a ponto dela, tá? Então, plt ponto plot min, aí aqui vai o ytest vírgula maxytest, fecho parênteses vírgula minytest maxytest de novo, red linewidth. Bom, a gente tem aqui o plt ponto plot que vai criar essa reta da nossa regressão. Então, aqui a gente tem de primeiro argumento o ponto de início da reta e depois o ponto de final da reta. Um ponto tem um ponto no x aqui, um valor do x, um valor do y para ter um ponto. Depois um valor do x, um do y para ter outro ponto e criar essa reta. Aqui no color, a gente está falando que essa reta vai ser da cor vermelha. e o line width, que é o tamanho da espessura da linha. A gente definiu como dois. Vocês podem modificar isso aqui, modificar a cor da reta, podem ficar à vontade. Bom, por fim, para visualizar, a gente coloca o plt ponto show. E vamos executar para ver como que vai ficar. Aqui.

00:00:03:22 - 00:00:04:51
Meu gráfico ficou muito grande pra ver, vou diminuir aqui um pouquinho, tá gente? Ó, de 8 vou colocar 4 aqui. Vamos ver se vai ficar melhor. Melhorzinho, porque aí dá pra poder ver aqui numa tela. Bom, então nesse gráfico temos que a linha vermelha representa o nosso modelo de regressão linear, ou seja, os valores preditos. As bolinhas azuis são os valores reais dos salários. Podemos ver que abaixo de 10 mil reais aqui de salários, temos uma concentração de bolinhas, ou seja, muitos salários estão nessa faixa. Analisando visualmente a performance do nosso modelo em relação aos valores reais, podemos ver que até aproximadamente 20 mil aqui, até aproximadamente 20 mil reais de salário, temos muitas bolinhas azuis, que são os valores reais, próximos dessa linha vermelha, que é a do modelo. Existem bolinhas acima da linha e abaixo, mas estão próximas da linha em si, seguindo a mesma inclinação. Porém, acima de 20 mil reais, as bolinhas estão bem distantes da linha, correto? Elas seguem uma tendência mais achatada, ao invés de subir aqui, elas dão essa achatada, que não casa com a inclinação da regressão que criamos. Isso significa que ou podemos tentar no modelo com inclinação menor, ou realmente regressão linear não é um modelo adequado para o nosso problema.

00:00:04:51 - 00:00:06:27
Porém, como no nosso caso o objetivo real é fazer uma análise das relações com o salário, e na maioria dos casos conseguimos uma predição próxima, usaremos esse modelo assim mesmo. Nós podemos ver também quais atributos tiveram mais peso positivo ou negativo para o resultado do modelo. Em modelo de regressão, os coeficientes ou pesos eram aqueles multiplicadores que a gente viu na fórmula de regressão. Bom, para ficar mais claro, então, nesse exemplo aqui de 4 açúcares, 2 leites, 2 ovos, 8 farinhas, 1 óleo, a gente vê que farinha tem esse 8 aqui, que seria esse multiplicador. Então, farinha seria o mais importante nessa equação, teve um peso maior nessa equação do que óleo, por exemplo, que tem o valor 1 aqui, esse multiplicador 1. Compreender esses coeficientes pode ajudar a interpretar o modelo e entender quais características são mais significativas. Bom, vamos voltar aqui e primeiro vamos pegar os nomes das colunas de atributos e atribuir a uma variável. Vamos lá. Aqui mesmo, nessa linha de código, vamos chamar de nomes atributos igual a xtrain, que vai ser o nome das colunas, que são os nossos atributos. Xtrain, se a gente tem mais de um train, o nome de xtrain. Se a gente ver, está aqui o nome dos nossos atributos, das nossas colunas. Vamos criar um DataFrame com pandas, uma tabela com pandas, pd ponto DataFrame. Esse DataFrame, ou seja, essa tabela vai ter como valores os coeficientes do modelo que pode ser acessado assim: model ponto coef underline...

00:00:06:27 - 00:00:07:59
Isso aqui são os coeficientes do modelo que a gente treinou. O segundo parâmetro é para dizer qual o nome dessa coluna que a gente acabou de criar. Vamos chamar de coeficientes, assim. A gente pegou o model coef, a gente pegou os coeficientes, e agora a gente está dando um nome para essa coluna. Então, columns igual a coeficientes. Beleza? Por último, vamos passar qual vai ser o nome das linhas, dos indexes, que no caso serão os nossos atributos. Então aqui coloco vírgula, coloco index nomes atributos, ok? Perfeito, se a gente executar aqui, a gente já vai ter a tabela. Mas melhor a gente armazenar esse resultado numa variável, né? Vamos chamar de coefs. Coefs, só para não precisar escrever coeficiente. Aqui: coefs. Vamos ordenar essa coluna de coeficientes para a gente conseguir ver melhor o que está influenciando mais ou o que está influenciando menos. Vamos fazer isso usando a função de sort values, que a gente já viu aí em módulos passados. coefs ponto sort values, para poder ordenar, e a gente coloca através de qual coluna que vai ordenar: através da coluna de coeficientes, ok?

00:00:07:59 - 00:00:09:34
E aí a gente vai colocar ascending igual a false, ok? Bom, vamos executar primeiro, antes de salvar direto para ver se está certinho? Beleza. Está do que mais influenciou para o que menos influenciou, do que teve um peso maior para o que teve peso menor. Vamos salvar esse resultado no coefs. Para ficar melhor ainda, a gente pode plotar um gráfico de barra horizontal, que aí a gente vai ver realmente visualmente o que está influenciando mais ou menos. A gente pode fazer isso: coef ponto plot ponto bar, que é de barra na horizontal, e aí a gente coloca um fig size aqui dentro. De novo, vai ser um chute esse fig size, vamos ver se vai dar certo aqui o valor. Vamos executar. Ó, eu acho até que ficou bom, hein, que dá para a gente poder ver. Vamos adicionar uma linha aqui na vertical, no eixo zero ali, no ponto zero, só para a gente poder ter essa referência para mostrar onde os coeficientes são positivos e onde eles são negativos, tá?

00:00:09:34 - 00:00:11:06
Então, vamos colocar aqui, ó, embaixo, plt ponto ax vline x zero e o color vamos colocar color red também, de novo. Vamos executar. Agora sim, está essa linha bem vermelha aqui no meio. Assim temos que todas as barras antes do zero têm um peso negativo para a predição do salário, ou seja, diminuindo o valor de salário e todas as barras acima de zero têm um peso positivo, aumentando o valor de salário. Vejam, a gente tem aqui o que mais influencia para o salário ser alto é se a pessoa é gestora, aumentando o valor desse salário, né? Depois, se a pessoa é nível sênio, depois é pelo tempo de experiência, o que faz sentido, né? Se a gente olhar no contexto, faz sentido uma pessoa gestora, e quanto mais tempo a pessoa tem de experiência, ela vai escalando aí nesses níveis de senioridade também, né? Começa júnior, pleno, sênio, pessoa gestora. Bom gente, vamos olhar aqui agora o que pesou para o lado negativo, quais foram os atributos que tiveram peso para colocar o salário mais para baixo. A gente teve aqui o que mais pesou foi pessoas que moravam na região do Sul. Depois a gente teve a insatisfação, mas aí a gente tem uma questão, não é? Não é porque a insatisfação que coloca o salário pra baixo, é porque essas pessoas já tinham um salário abaixo e elas estavam insatisfeitas com essa questão, né? Bom, a gente tem também a outra questão aqui de pessoas que moram no Nordeste, depois a gente tem, quase com o mesmo peso, a questão do não branca.

00:00:11:06 - 00:00:12:06
Então, acho que até, ressaltando de novo o que todas as nossas análises fizeram até agora, mostrando isso aqui, realmente escancada, que realmente pessoas, por exemplo, não brancas, têm um peso aí na questão do salário, elas ganham menos, um peso negativo para o valor do salário delas. Bom, dêem uma olhada assim, passem pelos pontos, conversem entre vocês, troquem uma ideia. Tem uma questão aqui de alguns setores que proporcionam aí um peso positivo, outros setores que proporcionam um peso negativo. Eu acho super válido assim. Tem muita coisa legal que pode ser discutida, tá? Um ponto aqui que eu acabei de olhar: olha só, o gênero masculino aqui, o peso positivo também que ele tem pro valor do salário ser alto no caso, né? Gente, uma coisa importante pra gente finalizar é que esse modelo nós estamos utilizando pra análise, tá?

00:00:12:06 - 00:00:13:05
Pra entender no mundo real de hoje o que influencia o salário. Por isso estamos olhando aqui os coeficientes e o que pesa pra negativo e positivo, a gente tá conversando, a gente tá criando aí algumas reflexões, colocando na sociedade e tal. Se um modelo para estimar o salário de uma pessoa para ser usado na vida real, por exemplo, o salário da equipe da PrograMaria vai ser estimado por um modelo, a gente teria que ter muito cuidado com ética para não reproduzir os vieses do mundo real. Caso a gente fosse fazer um modelo para estimar salários, aí a gente precisaria balancear, ou seja, equilibrar o modelo em relação à etnia, gênero, região e etc para não desfavorecer pessoas por essas variáveis. Esse assunto de ética na área de TI, de forma geral, é muito interessante. E a gente está deixando para vocês os materiais complementares, o link de uma matéria bem legal sobre viés lá no site da PrograMaria. Bom, gente, a gente chega aqui no final do nosso curso de análise de dados. Mas veja, a gente passou por muitas etapas, a gente fez muita coisa.

00:00:13:05 - 00:00:14:03
Foi um processo bem longo, né? E o que eu queria ressaltar para vocês é o nosso projeto, a nossa base que a gente trabalhou aqui. A PrograMaria tem como foco trazer essa diversidade para a área de dados. Então a gente trabalhou com essa questão da diversidade. Eu sei que olhando aqui o resultado, a gente fica: poxa, nossa, mas aqui está nesse peso negativo. Mas a gente está aqui, e se você está aqui fazendo esse curso, é porque você está querendo mudar isso, é porque você está querendo ser nessa parte positiva, a gente está querendo sair desse negativo e vir para esse positivo aqui. Então, a gente pegou a planilha, a gente começou com Excel, depois a gente foi para o Python, a gente passou por várias ferramentas, a gente fez um dashboard incrível, poxa. Sério, pega a família, coloca na sala, sabe, apresenta no PowerPoint assim: olha que bacana isso aqui que eu fiz. Coloca todo mundo para discutir e tudo mais. A gente trabalhou vários gráficos bacanas, a gente falou sobre banco de dados, cara, banco de dados, muito legal.

00:00:14:03 - 00:00:15:01
A gente fecha agora com esse módulo de aprendizado de máquina, que eu sei que é um módulo mais denso, um pouco mais complexo, mas a gente chega aqui num resultado muito bacana que a gente pode criar mais discussões. Lembrando que o nosso modelo tem poucos dados, né? A gente não pode falar cem por cento: nossa, é isso aqui que acontece. Mas a gente tem um conhecimento de sociedade, de contexto, que a gente sabe que algumas coisas aqui estão mesmo refletindo a nossa sociedade, né? Estão mesmo refletindo como as coisas são. Mas é isso, gente. Ah, uma coisa legal também: a gente fez um projeto passo a passo juntes, mas eu quero que vocês façam um projeto de vocês, sabe? Essa pesquisa da Data Hackers é feita todo ano. A gente usou a planilha de 2022. Dá uma olhada se já lançou a planilha de 2023 e atualiza aí a análise, faz a análise de novo com os dados de 2023 e coloca, vê se teve uma mudança de 2022 para 2023.

00:00:15:01 - 00:00:15:52
Eu acho que é um projeto maravilhoso. E vocês que fizeram esse curso e chegaram até aqui, vocês estão começando na área de dados, tem muita coisa para poder ser explorada, tá? Tem muita coisa que vocês podem fazer. A PrograMaria tem vários eventos aí para vocês se engajarem na comunidade, trocarem ideia entre vocês, trocarem ideia com pessoas que já estão trabalhando na área também. Procurar referência, gente, é uma coisa muito importante também. E é isso, eu estou muito feliz de a gente ter chegado até aqui, fico muito orgulhosa que a gente chegou nessas conclusões, que a gente chegou até aqui nesse ponto de análise, espero que vocês também estejam, e estou aguardando vocês compartilharem os projetos de vocês, porque eu vou estar olhando todos, hein? Quero admirar cada projetinho, cada detalhe que vocês compartilharem. E é isso, muito obrigada por terem participado aí.