1 Introdução

O domínio Inteligência Artificial (IA) na computação esta sendo fundamental na atual transformação digital que nossa sociedade esta passando e tornou-se um tema dominante na economia global. IA é um programa de computador que imita a inteligência humana de alguma forma e na grande maioria das vezes, esses programas usam técnicas de ciência de dados. Machine Learnig (ML) é uma das capacidades de IA e vem se destacando no mundo academico e corporativo como uma ferramenta que ajuda a resolver diversos problemas que as empresas enfrentam independentes do segmento de atuação e das áreas envolvidas. Nesta situação ML tem sido empregada das mais diferentes formas, especialmente na ciência de dados onde utiliza a modelagem estatística e funções como base para a realização de estimativas, sistemas de classificação, sistemas de recomendação, encontrar padrões e fazer detecção de anomalias entre outras.

2 Machine Learning

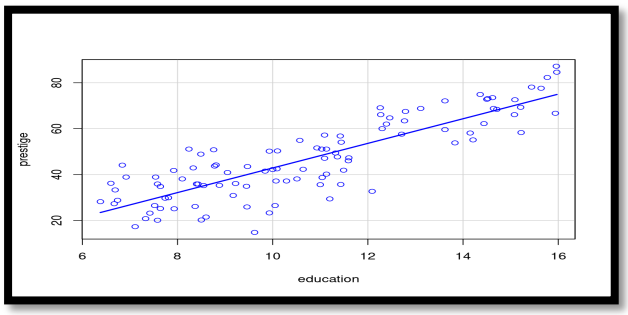
O aprendizado de máquina (ML) é um campo de investigação dedicado a entender e construir métodos como 'aprender', ou seja, métodos que aproveitam os dados para melhorar o desempenho em algum conjunto de tarefas. Isso é visto como parte da inteligência artificial. Algoritmos de aprendizado de máquina constroem um modelo baseado em dados de amostra, conhecidos como dados de treinamento, a fim de fazer previsões ou decisões sem ser explicitamente programado para isso.
Para resolver os diferentes problemas que enfrentam, os algoritmos de ML utilizam dois tipos de aprendizado: o supervisionado e o não supervisionado na elaboração de diversos modelos para indetermináveis fins.
O tipo de aprendizado não supervisionado é apropriado para os dados que não estão estruturados não possuem uma classe conhecidas e são resolvidos por meio da clusterização, associação ou redução das dimensão das variaveis.
O tipo de aprendizado supervisionado é onde temos uma variável target ou uma classe que é identificada no seu conjunto de dados e são resolvidos por meio da das técnicas estatísticas de regressão logística fazendo a classificação nos modelos e pela regressão linear realizando a predição nos modelos.

3 Regressão Linear

Temos dois tipos de técnicas de regressão: a simples e a múltipla. A regressão linear simples é uma espécie de modelo na estatística cujo objetivo é indicar qual será o comportamento de uma variável dependente (Y) como uma função que contenha uma ou mais variáveis independentes (X). Na regressão linear simples, a relação entre duas variáveis pode ser representada por uma linha reta, criando uma relação direta de causa e efeito. Assim, será possível prever os valores de uma variável dependente com base nos resultados da variável independente, é estabelecida uma equação:

$$Y = \beta_0 + \beta_1 X + E$$

A demonstração gráfica da regressão linear simples é mostrada abaixo com relacionamento das variaveis prestigio(Y) e educação(X).



Muitas vezes uma única variável explicativa (preditora) não será capaz de explicar tudo a respeito da variável resposta. Se em vez de uma, forem incorporadas várias variáveis independentes, passa-se a ter uma análise de regressão linear múltipla. Nesse caso, a equação estabelecida é:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + E$$

3.1 Métricas da Regressão Linear

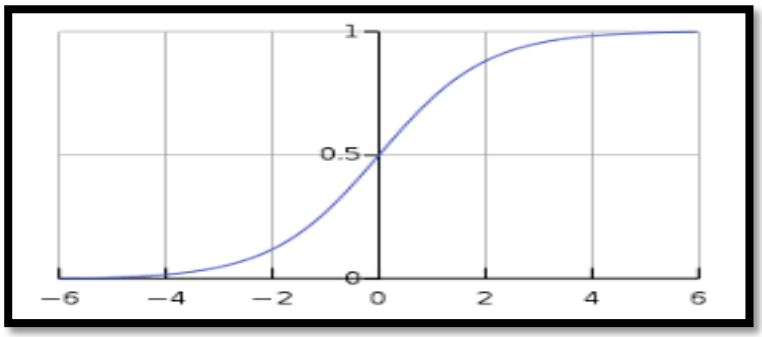
A métrica de avaliação para cada caso deve ser capaz de informar adequadamente sobre os capacidade de capturar com precisão as relações nos dados. A seleção da métrica deve ser determinado pelo tipo de problema com o objetivo da avaliação do resultados apresentados pelos modelos propostos.
A principais metricas para avaliação da regressão linear são o MSPE, RMSE R Square, Ajusted R Square, onde:
MSPE: O erro percentual absoluto médio, também conhecido como desvio percentual absoluto médio, é uma medida de precisão de previsão de um método de previsão em estatística.
RMSE: O desvio médio quadrático ou erro quadrático médio é uma medida freqüentemente usada das diferenças entre os valores previstos por um modelo ou estimador e os valores observados.
R Square (R^2): O coeficiente de determinação, também chamado de R^2, é uma medida de ajuste de um modelo estatístico linear generalizado, como a regressão linear simples ou múltipla, aos valores observados de uma variável aleatória. O R^2 varia entre 0 e 1, por vezes sendo expresso em termos percentuais.
Ajusted R^2 Square: é uma medida corrigida de qualidade de ajuste (precisão do modelo) para modelos lineares. Ele identifica a porcentagem de variação no campo de destino que é explicada pela entrada ou entradas. O R^2 ajustado tende a estimar com otimismo o ajuste da regressão linear.

4 Regressão Logística

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias. A regressão logística é usada para determinar a probabilidade de um evento acontecer. Ele mostra a relação entre os recursos e, em seguida, calcula a probabilidade de um determinado resultado. Podendo ser aplicada na previsão de risco na area tributaria – calcular a probabilidade do contribuinte ser inadimplente o adimplente após o parcelamento de tributos, utilizado para classificar se a empresa encontra-se no grupo de empresas solvente ou insolvente. Matias e determinar quais características levam as empresas adotarem um modelo de indicadores de gestão, ex. *balanced scorecard*.
Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente *Y* assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de *p* variaveis independentes *X1* , *X2* , ... , *Xp,* o modelo de regressão logística pode ser escrito da seguinte forma:

$$P(Y=1)=\frac{1}{1+e^{-g(x)}}$$

A demonstração gráfica da regressão logística é chamada de Sigmoid em forma de S, conforme demostrado abaixo:



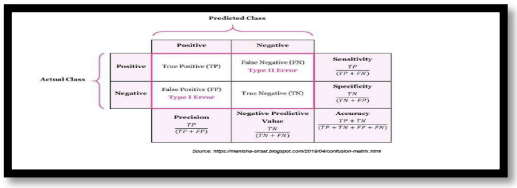
A intuição por trás de regressão logística é bastante simples: em vez de acharmos a reta que melhor se ajusta aos dados, vamos achar uma curva em formato de 'S' que melhor se ajusta aos dados.

4.1 Métricas da Regressão Logística

A métrica de avaliação para cada caso deve ser capaz de informar adequadamente sobre os capacidade de capturar com precisão as relações nos dados. A seleção da métrica deve ser determinado pelo tipo de problema com o objetivo da avaliação do resultados apresentados pelos modelos propostos.

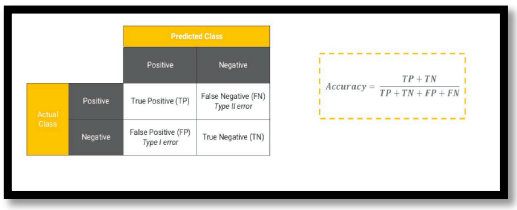
A principais metricas para avaliação da regressão logística são extraídas da Matriz da confusão e são a Acurácia, Precisão/Recall, F1 e AUC (área abaixo da curva ROC. Uma matriz de confusão é uma tabela que indica os erros e acertos do seu modelo, comparando com o resultado esperado (ou etiquetas/*labels*).

Matriz da Confusão

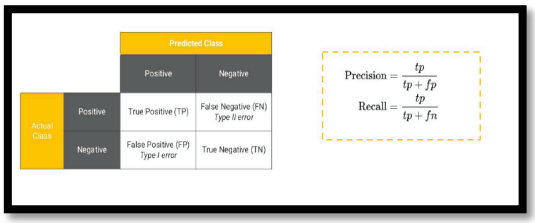


H0= hipótese nula H1= hipótese alternativa
Erro tipo 1= Falso Positivo Erro tipo 2= Falso Negativo

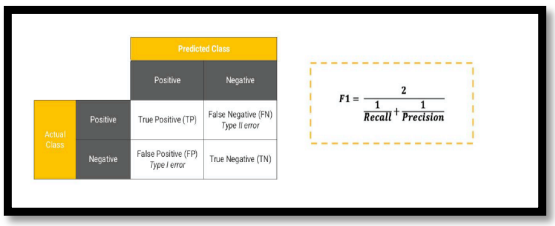
Acurácia indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;



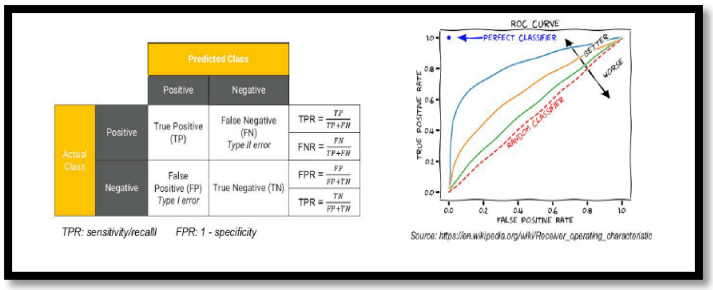
Precisão: dentre todas as classificações de classe Positivo **que o modelo fez**, quantas estão corretas; ***Recall*/Sensibilidade:** dentre todas as situações de classe Positivo **como valor esperado**, quantas estão corretas;



O **F1-Score** é simplesmente uma maneira de observar somente 1 métrica ao invés de duas (precisão e recall) em alguma situação. É uma média harmônica entre as duas, que está muito mais próxima dos menores valores do que uma média aritmética simples.



Curvas ROC (receiver operator characteristic curve) são uma forma de representar a relação, normalmente antagónica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo, ao longo de um contínuo de valores de ponto de corte.



Cada métrica tem suas peculiaridades que devem ser levadas em consideração na escolha de como o modelo de classificação será avaliado. Não se deve pensar em uma como melhor ou pior que a outra de maneira geral, e sim deve-se analisar o problema e escolher a/as que melhor se adapta(m).

5 Conclusão

A regressão linear busca a correlação entre duas variaveis numéricas e a regressão logística pode ter variaveis numéricas ou categóricas. A regressão logística é utilizada quando a resposta ao seu problema pode ser categorizada. ex: quais funcionários tem maior chance de sair da empresa? Neste caso esta pergunta é de resposta binária...classificativa, ou o funcionário sai ou fica e o modelo separará este dois blocos e você pode avaliar por meio da matriz de confusão

A principal diferença da regressão logística para a regressão linear é que a variável dependente na regressão logística é categórica, e o modelo de regressão logística é o modelo mais importante para dados de resposta categórica.

Então o emprego de uma ou outra técnica fica a cargo do problema de negocio a ser resolvido pois ambas as técnicas são robustas em cumprir os seus objetivos sejam na previsão ou estimativa (regressão linear) ou de classificação (regressão logística).

6 Bibliografia

BIDO, D. S. *et al.* Indicadores Formativos na Modelagem em Equações Estruturais com Estimação via PLS-PM: Como Lidar com a Multicolinearidade Entre Eles? ***EnEPQ***, p. 1–16, 2009.

BIDO; SILVA, D. SmartPLS 3: especificação, estimação, avaliação e relato. ***Administração: Ensino e Pesquisa***, v. 20, n. 2, p. 488–536, 2019.

HAIR *et al*. *Analise Multivariada de dados*. Ebook, ***Artmed Editora***. 5, 1998.

HAIR, J. *et al*. A primer on partial least squares structural equation modeling (PLS-SEM). ***International Journal of Research & Method in Education***, v. 2, n. 2, p. 220–221, 2015.

HAIR, J. F. *et al.* Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. ***European Business Review***, v. 26, n. 2, p. 106–121, 2014.

HAIR, J. F.; RINGLE, C. M.; SARSTEDT, M. PLS-SEM: Indeed a silver bullet. ***Journal of Marketing Theory and Practice*****,** v. 19, n. 2, p. 139–152, 2011.