**ML: Credit Scoring**

A Cadeia de crédito é composta por 4 pilares: **prospecção**, **concessão**, **gestão de risco** e **recuperação**.

Os principais participantes da cadeia de crédito são os **poupadores**, os **tomadores** e as **instituições financeiras** que intermediam essa relação.

Quando um tomador solicita um crédito, a instituição financeira avalia os dados cadastrais e financeiros através de um **modelo de decisão** chamado **Credit Scoring**.

O Credit Scoring **estima a probabilidade** de um tomador se tornar inadimplente, classificando-os como **adimplente** ou **inadimplente**. Essa probabilidade está relacionada ao **risco de crédito**.

O risco de crédito é a probabilidade de a instituição financeira não receber o valor emprestado de volta no prazo acordado. Parte dos juros cobrados está embutida nesse risco.

Portanto, a **concessão de crédito** **está intrinsecamente** **relacionada ao risco**, que é um fator crucial nas decisões das instituições financeiras.

**O que é o Credit Scoring?**

O credit scoring é um **modelo estatístico multivariado** que estima a probabilidade de um cliente ser adimplente ou inadimplente. Do ponto de vista do banco, quando um cliente solicita um crédito, o banco analisa essa solicitação usando um modelo de credit scoring.

Alguns algoritmos comumentes usados para credit scoring são a **regressão logística**, **Random forest**, **Naive Bayes** e **support vector machine**.

A regressão logística é uma boa opção porque tem alto comprometimento com a explicabilidade, permitindo acompanhar bem as variáveis ao longo do tempo, e exige pouco poder computacional, o que é importante dado o tamanho das bases de dados de bancos.

* Como funciona uma solicitação de crédito;
* O que é credit scoring;
* Calcular o total de linhas e colunas;
* Identificar dados duplicados;
* Identificar dados nulos.

Regressão Logística

Base de dados para treino e teste

É importante ter uma variável resposta binária e variáveis independentes conhecidas em nossa base de dados. A variável resposta, chamada default, é analisada junto com variáveis categóricas e discretas.

O primeiro passo é separar a base de dados em duas partes: X, que contém as variáveis independentes, e Y, que é a variável resposta. Utilizamos o método ‘drop’ para criar a base X, excluindo a coluna default.

Em seguida, dividimos as bases X e Y em conjuntos de treino e teste usando a função ‘train\_test\_split’ do Scikit-learn. Definimos o tamanho do conjunto de teste (test\_size) e um estado aleatório (Random\_state) para garantir que a divisão seja replicável. Por fim, verificamos o número de elementos em cada conjunto, confirmando que a divisão foi realizada.

Transformação de categorias

Aprendemos sobre a transformação de variáveis categóricas em valores numéricos para que possamos utilizar a regressão logística. O processo envolve os seguintes passos:

**Identificação de Variáveis Categóricas:** Primeiro identificamos quais colunas da nossa base de dados contêm strings que representam categorias, como a coluna “conta\_corrente”.

**Verificação de Valores Únicos:** Utilizamos o método ‘unique()’ para descobrir quais são os diferentes valores que essa variável pode assumir.

**Criação de um Dicionário:** Em seguida, criamos um dicionário onde cada categoria é mapeada para um número.

**Aplicação do Mapeamento**: Usamos o método ‘map()’ para substituir as strings na coluna pela representação numérica correspondente, transformando assim a variável categórica em uma variável numérica.

Após fazer a transformação em outras categorias, foi realizado o train test split para dividir os dados em conjuntos de treino e teste, com uma semente definida (SEED = 77) para garantir a reprodutibilidade. Após a divisão, foi impresso o número de elementos em cada base, mostrando que a base de treino tinha 700 elementos e a base de teste, 300.

Por fim, criei o modelo de regressão logística e ajustei aos dados de treino, e a acurácia do modelo foi calculada, resultando em um valor de 0.74, que representa a precisão do modelo em prever corretamente os dados de treino. Essa é a primeira métrica estatística a ser avaliada;