Conversas: Serasa – Covariate Drift

FELIPE LENO: AI2 - Advanced Institute for Artificial Intelligence

Motivação

- Score de crédito

Score de crédito

Score de crédito

Score de crédito

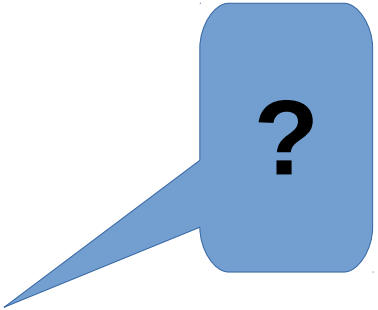
Score de crédito

- Cliente adimplente

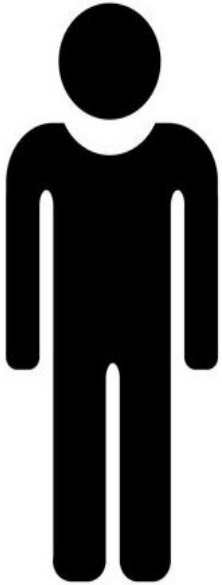
Score de crédito

Score de crédito

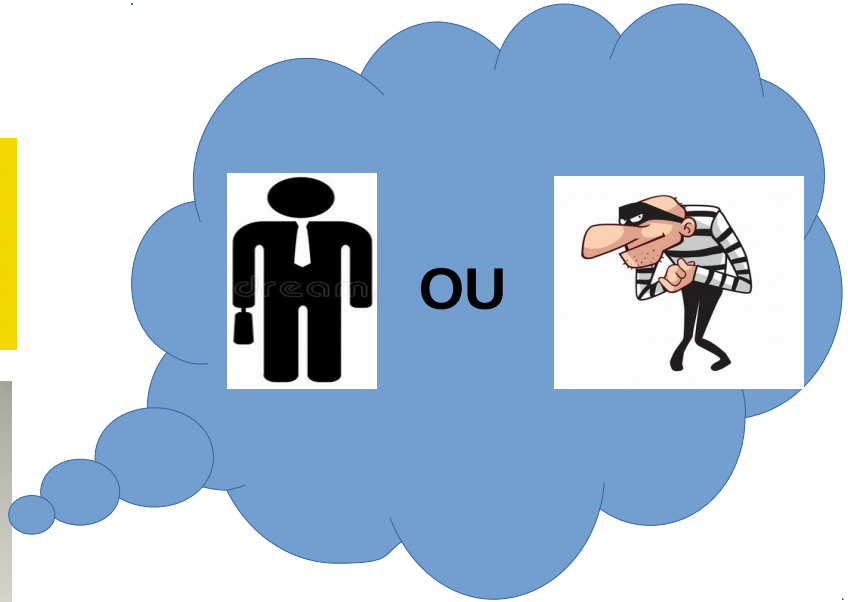
- Cliente inadimplente

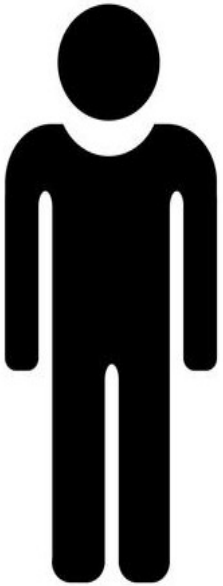


Score de crédito



Score de crédito





Como saber se um potencial cliente é confiável?

- Histórico de inadimplência
- Histórico de serviços contratados
- Renda e outras informações sócio-econômicas

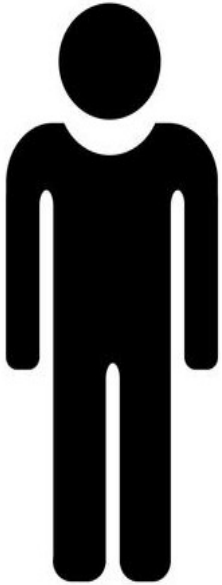
Como saber se um potencial cliente é confiável?

- Histórico do cliente é geralmente limitado ou até mesmo completamente indisponível (se é um cliente novo)

O que fazer?

Score de crédito

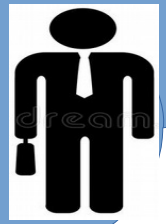
650



Score de crédito

Score Baixo =

Score Alto =

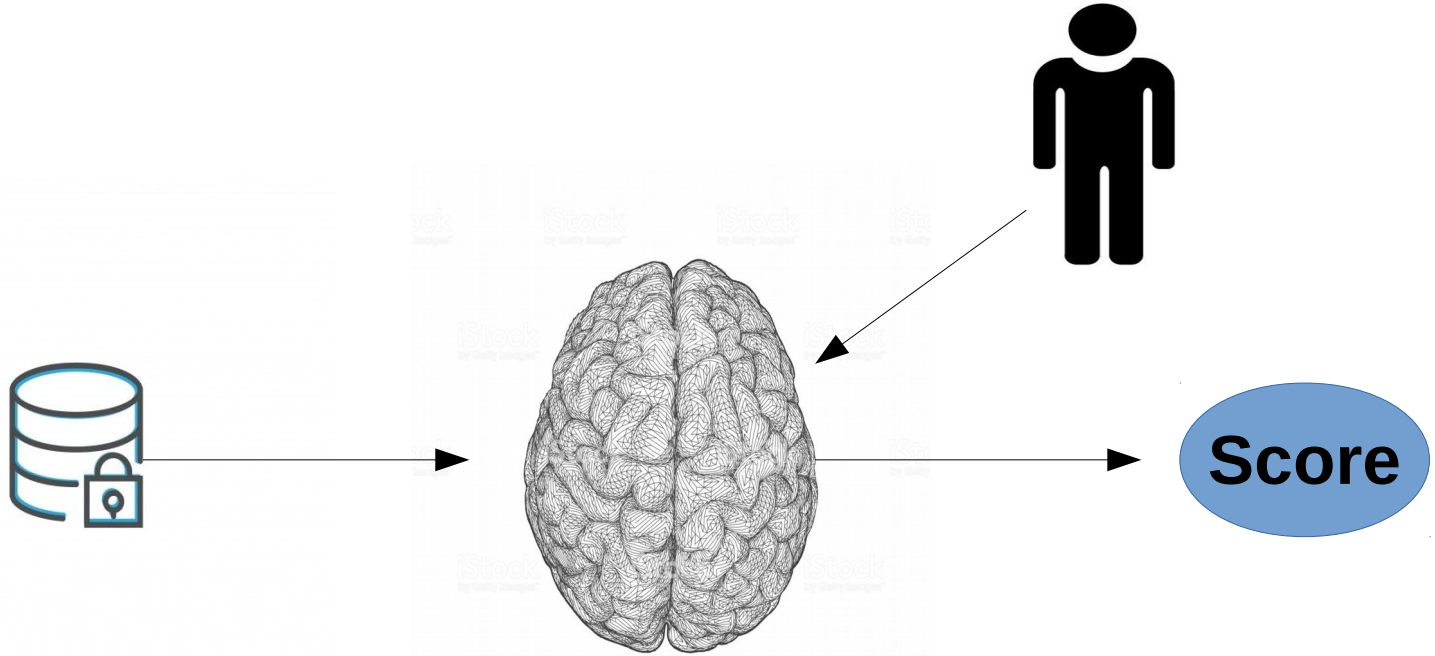


Como o Serasa faz isso?

Como o Serasa faz isso?

- Se o cliente é novo, generaliza clientes similares

Problema de Classificação

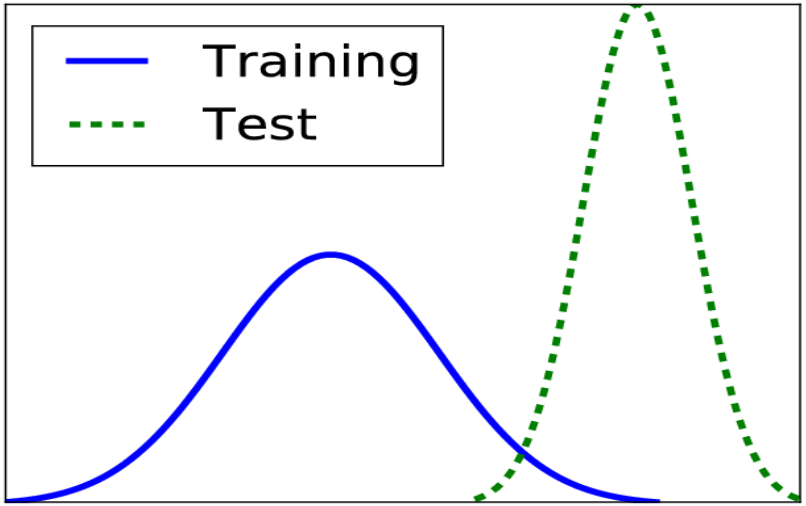


Problema de Classificação

- Probabilidade de ser inadimplente é transformada no score

Desafio de Pesquisa

- O que é Covariate Shift?



Implicações no problema real

- Dados coletados são “velhos”
  - Situações pessoais mudam rapida e constantemente
  - Situações do mercado mudam raramente mas podem ter impacto profundo e rápido

Desafio 1

- Como identificar “quando” e “como” a distribuição dos dados muda
  - Identificar quando é necessário retreinar o modelo
  - Reamostrar o conjunto de treinamento para otimizar performance

Desafio 2

- Migrando para novas regiões

Muitos dados

Fase 1: Nenhum dado
Fase 2: Poucos dados

Desafio 2

- Como reutilizar dados brasileiros no contexto “sul-africano”, para construir um modelo rapidamente mesmo sem ter tido tempo de montar uma base de dados?

O que faremos?

- Trabalhar em algum dos desafios **utilizando bases simuladas**

- Ainda trabalhando na revisao da literatura, mas acredito que reusar a base em novas regiões com poucos dados é mais fácil de simular e avaliar em outras bases.