

Turma 8DTS

Data Science & Artificial Intelligence

STATISTICS WITH R

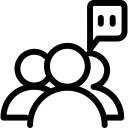
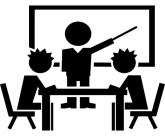
//adelaidealves

## Introdução



**Do** **pensamento** **à** **ação** - Escolha de organização

O case da fintech **Quantum** **Finance**



A Quantum Finance é uma nova fintech, que está entrando no mercado para concorrer com grandes players.

Neste case veremos como aplicar a ciência de dados e Artificial Intelligence (AI) a para a concepção e expansão da empresa no mercado nacional e internacional, em diversos segmentos.

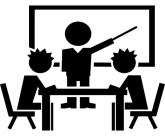
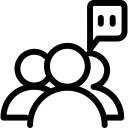


## Introdução



### Do pensamento à ação - Quantum Finance

O case da fintech **Quantum** **Finance**



A Quantum Finance é uma nova fintech, que está entrando no mercado para concorrer com grandes players.

Neste case veremos como aplicar a ciência de dados e Artificial Intelligence (AI) a para a concepção e expansão da empresa no mercado nacional e internacional, em diversos segmentos.

## Direcionadores da organização

Os steakholders e founders da Quantum Finance estão empolgados para começar o novo negócio da forma mais automatizada e inteligente possível. Porém estão em dúvida sobre quais tipos de soluções podem ser resolvidas por AI, onde devem ou não utilizá-la e com quais tecnologias.

Então, convidam a consultoria SmartMind Analytics para estruturar a empresa e, utilizando a metodologia ágil **Strategic** **Thinking** **for** **Business**, irão definir seus principais direcionadores, partindo de sua proposta de valor.

PRODUTO

CRÉDITO AO CONSUMIDOR



A ***QuantumFinance*** está acompanhando um crescimento de inadimplência entre seus clientes e solicita a consultoria para desenvolver uma análise com base na sua carteira atual de clientes.

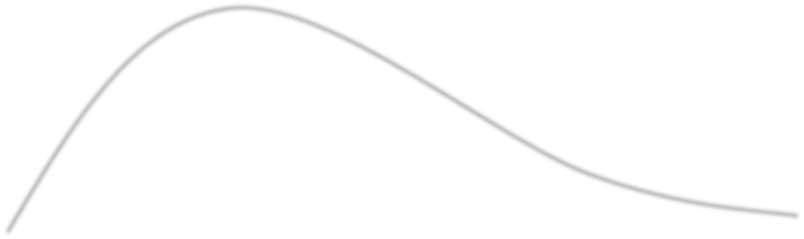
Para que a ***QuantumFinance*** tome decisões mais precisas sobre concessões de crédito, ela precisa aprimorar seu modelo de crédito.

Com o objeto de trazer novos clientes com o perfil de baixo risco de crédito desenvolva um modelo de Credit Scoring.

Desafio: Desenvolver o modelo preditivo mediante uso do valor target disponível na base de dados “**Base\_ScoreCredito\_QuantumFinance.csv**” e criar um simulador do modelo para os analistas de créditos e gerentes de conta.

### Uso dos Modelos no Ciclo do Cliente

* **Credit** **Scoring**
* **Score** **de** **Risco** **(Pagamento)**
* **Modelo** **de** **inadimplência**



* **Valor** **do** **Cliente**
* **Modelos** **de** **Churn**
* **Detecção** **de** **Fraude**
* **.......**
* **Collection** **Score**
* . ...

# PROSPECÇÃO

**Aquisição** **Qualificada**

# MANUTENÇÃO

**Fidelização** **com** **Ampliação**

# RETENÇÃO

**Retenção** **com** **Valor**

# RECONQUISTA

**Reconquista** **Qualificada**

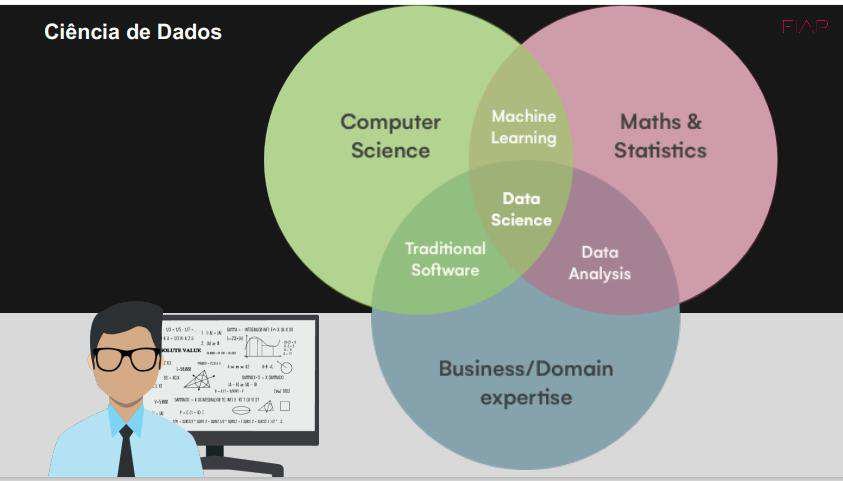
### Tipos de Modelos

* + **Modelo** **de** ***Credit*** ***Scoring***
    - Considera informações/dados do contrato (tempo de relacionamento recente);
    - Score de crédito de um novo cliente
  + **Modelo** **de** **Inadimplência**
    - Considera dados de utilização dos clientes;
    - Probabilidade do cliente vir a ser inadimplente.
  + **Modelo** **de** ***Behaviour*** ***Score***
    - Considera dados de utilização dos clientes; Dados comportamentais de pagamento
    - Score comportamental de crédito.
  + **Modelo** **de** **Cobrança**
    - Considera dados de utilização dos clientes e do mercado
    - Probabilidade do cliente pagar
  + **Modelo** **de** **Churn** **e** **fraude/anomalias/** **abusos**
    - Considera dados de utilização dos clientes e do mercado
    - Probabilidade do cliente cancelar a “conta”

**Arquivo** **disponibilizado**

## DATASET

|  |  |  |
| --- | --- | --- |
| **Variáveis** | **Descrição** | **Categorias** |
| id | ID |  |
| idade | Idade em anos |  |
| sexo | Sexo | F = Feminino |
| M=Masculino |
| estado\_civil | Estado civil | Solteiro |
| Casado |
| Divorciado |
| escola | Grau de Instrução | ensino fundam = até o ensino fundamental |
| ensino médio |
| Graduação |
| Doutorado |
| Qte\_dependentes | Quantidade de dependentes |  |
| tempo\_ultimoservico | Tempo de trabalho (meses) |  |
| trabalha | Trabalha | 0=Não |
| 1=Sim |
| vl\_salario\_mil | Sálario |  |
| reg\_moradia | Região da moradia | 1=Segmento geográfico A |
| 2=Segmento geográfico B |
| 3=Segmento geográfico C |
| 4=Segmento geográfico D |
| 5=Segmento geográfico E |
| 6=Segmento geográfico F |
| casa\_propria | Casa própria | 0=Não |
| 1=Sim |
| vl\_imovel\_em\_mil | Valor do imóvel (R$) |  |
| Qte\_cartoes | Quantidade de cartões de crédito |  |
| Qte\_carros | Quantidade de carros |  |
| SCORE\_CREDITO | Credit Scorging |  |



Fonte: Marcelo Manzano. 2022.1 - DTS Foundation.pdf

Produto: Crédito ao Consumidor

|  |  |  |
| --- | --- | --- |
| Tópico | O que é? | Como está? |
| Crédito ao consumidor | Crédito ao consumidor refere-se à prática de fornecer empréstimos a indivíduos para atender a despesas pessoais, como compras, educação ou necessidades diárias. Instituições financeiras, como bancos e cartões de crédito, oferecem diversas formas de crédito, como cartões de crédito, empréstimos pessoais e financiamentos, permitindo que os consumidores realizem transações e paguem em parcelas ao longo do tempo. | O consumidor brasileiro tem diminuído a procura por empréstimos. Dados do Indicador de Demanda dos Consumidores por Crédito, da Serasa Experian, apontam que em julho houve uma queda de 2,4% na comparação com junho. Já na comparação com julho de 2022, a retração foi de 10,9%. |
| Credit Scoring | O crédito score é uma medida numérica que avalia o histórico de crédito de um indivíduo, refletindo sua capacidade de cumprir obrigações financeiras. Calculado com base em informações como histórico de pagamento, dívidas pendentes, tempo de crédito e tipos de crédito utilizados, o crédito score é usado por instituições financeiras para avaliar o risco de emprestar a um consumidor. Quanto maior o score, maior a confiança nas habilidades de pagamento do indivíduo, facilitando o acesso a empréstimos e taxas de juros mais favoráveis. Por outro lado, um score mais baixo pode resultar em restrições no acesso ao crédito ou condições menos favoráveis de empréstimo. | 485 pontos  O Serasa score médio do brasileiro é de 485 pontos. Os números assustam um pouco. Quase um terço das pessoas têm alto risco de inadimplência. A nota média é a metade da máxima, fica bem no centro da faixa de médio risco. |
| Inadimplência | A inadimplência ocorre quando um devedor deixa de cumprir suas obrigações financeiras, como não pagar empréstimos, contas ou outras dívidas dentro do prazo acordado. Isso pode resultar em consequências negativas, como multas, juros acumulados e danos ao histórico de crédito do devedor. | Com dados de Novembro de 2023 cerca de 71,81 milhões de brasileiros em situação de inadimplência, a queda foi de 143,5 mil em relação ao mês anterior. |
| Endividamento da população | O endividamento da população refere-se à situação em que os indivíduos ou famílias acumulam dívidas que podem resultar em desafios financeiros. Isso acontece quando as obrigações financeiras, como empréstimos, financiamentos ou contas de crédito, excedem a capacidade de pagamento dos devedores. | A cada 10 famílias brasileiras, 8 têm dívidas. Essa tem sido a taxa média de endividamento das famílias, medida pela Pesquisa de Endividamento e Inadimplência do Consumidor (Peic), divulgada pela Confederação Nacional do Comércio de Bens, Serviços e Turismo (CNC).  Em abril de 2023, 78,3% dos núcleos familiares do país tinham dívidas. Em 2022, a média total foi de 77,9%, valor recorde desde que o levantamento começou a ser feito pela confederação, em 2011. |

## Estudo Estatístico – Quadro Conceitual

|  |  |
| --- | --- |
| COMPONENTES | DESCRIÇÃO |
| 1.Tema | Construção e avaliação de modelos estatísticos para prever o risco de crédito de indivíduos com base em variáveis financeiras e de crédito. A análise estatística pode incluir a identificação de variáveis relevantes, a seleção e implementação de métodos estatísticos, como regressão logística ou árvores de decisão, a avaliação da precisão do modelo e a interpretação dos resultados. O objetivo é desenvolver um modelo robusto que permita às instituições financeiras avaliar o risco de inadimplência e tomar decisões informadas sobre a concessão de crédito. A análise estatística no contexto de credit scoring é crucial para desenvolver modelos preditivos eficazes e compreender as relações entre as variáveis influentes. |
| 2. Problema | Ao desenvolver uma análise estatística de credit scoring, é crucial evitar problemas comuns, como overfitting, seleção de variáveis irrelevantes, viés de amostragem, multicolinearidade e falta de interpretabilidade. A busca por equilíbrio entre a complexidade do modelo e sua interpretabilidade é fundamental para explicar as decisões de crédito a stakeholders e garantir a confiabilidade do modelo na prática. |
| 3. Hipóteses conceituais | Ao realizar uma análise estatística de credit scoring, as hipóteses conceituais fundamentais incluem a expectativa de que variáveis financeiras relevantes, como histórico de pagamento e renda, possuam uma relação significativa com o risco de crédito. Assume-se a linearidade dessa relação, a estabilidade temporal das variáveis preditivas e o impacto diferenciado de determinadas características, como histórico de inadimplência e limites de crédito. A normalidade dos resíduos é considerada, assim como a ausência de colinearidade entre variáveis independentes. Além disso, há a hipótese de que o modelo proposto supere alternativas em termos de desempenho, medido por métricas de precisão e validação estatística. Essas hipóteses guiam o desenvolvimento e a validação do modelo de credit scoring, buscando uma compreensão robusta e preditiva do risco de crédito. |

Continuação ...

## Estudo Estatístico – Quadro Conceitual

Continuação ...

|  |  |
| --- | --- |
| COMPONENTES | DESCRIÇÃO |
| 4. Objetivo principal | Prever e avaliar o risco de crédito associado a indivíduos ou entidades, fornecendo uma ferramenta analítica para as instituições financeiras tomarem decisões informadas sobre a concessão de crédito. |
| 5. População de estudo | Composta pelos indivíduos ou entidades que estão sujeitos à análise de crédito, ou seja, os clientes da instituição e os novos clientes. |
| 6. Plano básico de análise | O plano básico para a análise estatística de credit scoring envolve as seguintes etapas:  Definição do Problema e Objetivos:  Identificar claramente o objetivo da análise, como prever o risco de inadimplência em solicitantes de crédito.  Coleta de Dados:  Reunir dados relevantes, como histórico de pagamento, renda, dívidas e outras variáveis financeiras, para construir o conjunto de dados.  Exploração e Limpeza de Dados:  Explorar e entender os dados, identificando valores ausentes, outliers e inconsistências. Realizar a limpeza e transformação dos dados conforme necessário.  Seleção de Variáveis:  Escolher as variáveis mais relevantes e preditivas para inclusão no modelo, considerando a teoria financeira e a análise exploratória.  Divisão do Conjunto de Dados:  Separar o conjunto de dados em dados de treinamento e dados de teste para avaliar o desempenho do modelo.  Construção do Modelo:  Escolher e implementar um método estatístico apropriado, como regressão logística ou árvores de decisão, para construir o modelo de credit scoring.  Ajuste do Modelo:  Calibrar e ajustar o modelo com base nos dados de treinamento para melhorar sua capacidade preditiva.  Validação do Modelo:  Avaliar o desempenho do modelo usando dados de teste independentes para verificar sua eficácia na previsão do risco de crédito.  Interpretação dos Resultados:  Interpretar os resultados do modelo, identificando a contribuição relativa de cada variável e avaliando a lógica das previsões.  Documentação e Comunicação:  Documentar o processo, as decisões tomadas e os resultados obtidos. Comunicar os resultados de forma clara para as partes interessadas.  Implementação e Monitoramento Contínuo:  Implementar o modelo em ambientes de produção e estabelecer procedimentos para monitoramento contínuo, ajustes e atualizações conforme necessário. |

## Estudo Estatístico – Quadro Conceitual

Continuação ...

|  |  |
| --- | --- |
| COMPONENTES | DESCRIÇÃO |
| 7. Técnica estatística | Para a implementação do modelo de credit scoring, optaremos pela utilização da regressão logística como técnica estatística principal. A escolha pela regressão logística se justifica pela natureza binária da variável de resposta associada ao risco de crédito, onde o objetivo é prever a probabilidade de inadimplência. Essa técnica é apropriada para lidar com variáveis dependentes dicotômicas, como bons e maus pagadores, e proporciona interpretações claras dos coeficientes, permitindo uma análise mais profunda do impacto de cada variável explicativa na probabilidade de ocorrência do evento de interesse. A regressão logística é reconhecida por sua eficácia em modelos de credit scoring, fornecendo resultados sólidos e interpretáveis para apoiar decisões de concessão de crédito de maneira precisa e confiável. |
| 8. Resultado principal da técnica estatística | O principal resultado da aplicação da regressão logística em um modelo de credit scoring é a obtenção de coeficientes estimados para cada variável independente, representando a contribuição relativa de cada variável na probabilidade de inadimplência. Esses coeficientes são cruciais para a interpretação do modelo, permitindo avaliar a direção e magnitude do impacto de cada fator considerado. Além disso, o modelo fornece uma equação logística que relaciona as variáveis explicativas à probabilidade estimada de ocorrência da inadimplência. A precisão e validade do modelo podem ser avaliadas por meio de métricas como a área sob a curva ROC (Receiver Operating Characteristic) e a taxa de classificação correta. Essas informações são essenciais para instituições financeiras tomarem decisões informadas sobre a concessão de crédito, identificando potenciais riscos e otimizando a gestão de carteiras. |

Entregáveis:

- Arquivo word com os resultados:

Preencher o quadro conceitual estatístico; Análise descritiva das variáveis;

Análise de correlação das variáveis; Análise de resíduos;

Acurácia e medidas de erros do modelo.

Etapa 1

### Entrega em 31 de janeiro

* Faça recomendações positivas e negativas sobre as variáveis do modelo:
  + As variáveis são suficientes para tomada de decisão?
  + As suposições do modelo de regressão linear múltipla foram atendidas?
* Script do modelo.

Etapa 2

**Entrega** **em** **10** **de** **Março**

