

CoderHouse -
Kauan Rios de Melo

PREDIÇÃO DE CHURN EM TELECOM COM MACHINE LEARNING

Um estudo preditivo com foco na retenção de clientes

TABELA DE CONTEUDO

- 1. Introdução**
- 2. Objetivo**
- 3. Fonte de dados**
- 4. EDA (Analise exploratória de dados)**
- 5. Modelagem e otimização**
- 6. Avaliação de desempenho**
- 7. Conclusões**
- 8. Referência**

INTRODUÇÃO

- Este estudo tem como foco analisar e prever o cancelamento de clientes (churn) em uma empresa de telecomunicações. Através de ferramentas de ciência de dados, buscamos entender os padrões que levam os clientes a abandonarem o serviço.

OBJETIVO DO PROJETO

Desenvolver um modelo preditivo de machine learning que identifique, com base no histórico de dados, quais clientes possuem maior probabilidade de cancelar o serviço nos próximos meses.

A pesquisa busca entender os principais fatores que influenciam o churn de clientes em uma empresa de telecomunicações ou seja, por que os clientes estão cancelando seus serviços. A ideia é construir um modelo de classificação que seja capaz de prever com razoável precisão quais clientes têm maior probabilidade de abandonar a empresa, permitindo ações de retenção mais eficazes.

FONTE DOS DADOS

Dataset publico da IBM Telco Customer Churn.

Disponível em: <https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/>

Critérios: variáveis demográficas, contratuais e comportamentais



POR QUE ESTUDAR O CHURN?



- A perda de clientes impacta diretamente o faturamento.
 - O custo de aquisição de um novo cliente é maior que manter um atual.
 - Comportamentos de saída muitas vezes seguem padrões previsíveis.
 - A identificação antecipada permite ações de retenção.
- 

INDICAÇÃO DA FONTE DO DATASET E CRITÉRIOS DE SELEÇÃO (DATA ACQUISITION)

O dataset foi disponibilizado publicamente pela IBM e é amplamente utilizado para fins educacionais e estudos de ciência de dados. Ele pode ser encontrado em:

- A perda de clientes impacta diretamente o faturamento.
- O custo de aquisição de um novo cliente é maior que manter um atual.
- Comportamentos de saída muitas vezes seguem padrões previsíveis.
- A identificação antecipada permite ações de retenção.

Critérios de seleção:

- Selecionado por conter variáveis demográficas, comportamentais e contratuais dos clientes.
- A variável de interesse (Churn) já está presente e bem definida.
- Apresenta bom volume de dados (~7 mil linhas), o que é adequado para treinamento e validação de modelos.

ETAPAS DO PROJETO

Perguntas e Objetivos da Pesquisa

Objetivo

- Coleta e limpeza dos dados
- Análise exploratória (EDA)
- Pré-processamento (normalização + dummies)
- PCA para visualização
- Treinamento do modelo (Regressão Logística)
- Avaliação com métricas (AUC, F1-score)

Perguntas

- Quais variáveis têm maior correlação com o churn?
- Existem perfis específicos de clientes mais propensos ao cancelamento?
- O modelo de regressão logística é eficaz para prever o churn?
- Quais métricas de desempenho podemos usar para avaliar a performance do modelo?

ANÁLISE DOS DADOS

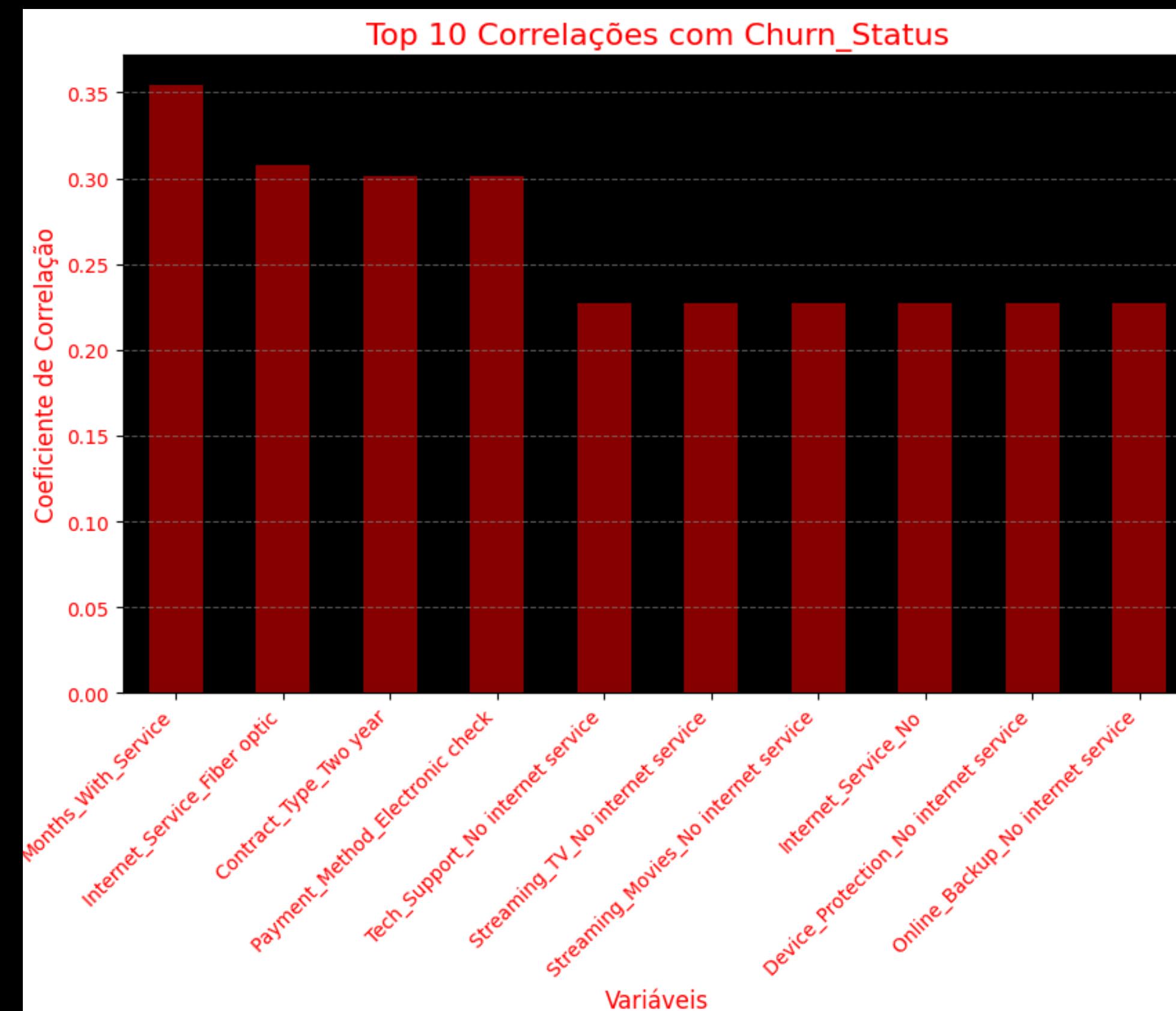
- A análise revelou correlações importantes entre tipo de contrato, tempo de serviço e valor mensal.
Clientes com contrato mensal e altos custos são mais propensos ao churn.



DISTRIBUIÇÃO DE CLIENTES POR STATÚS DE CHURN

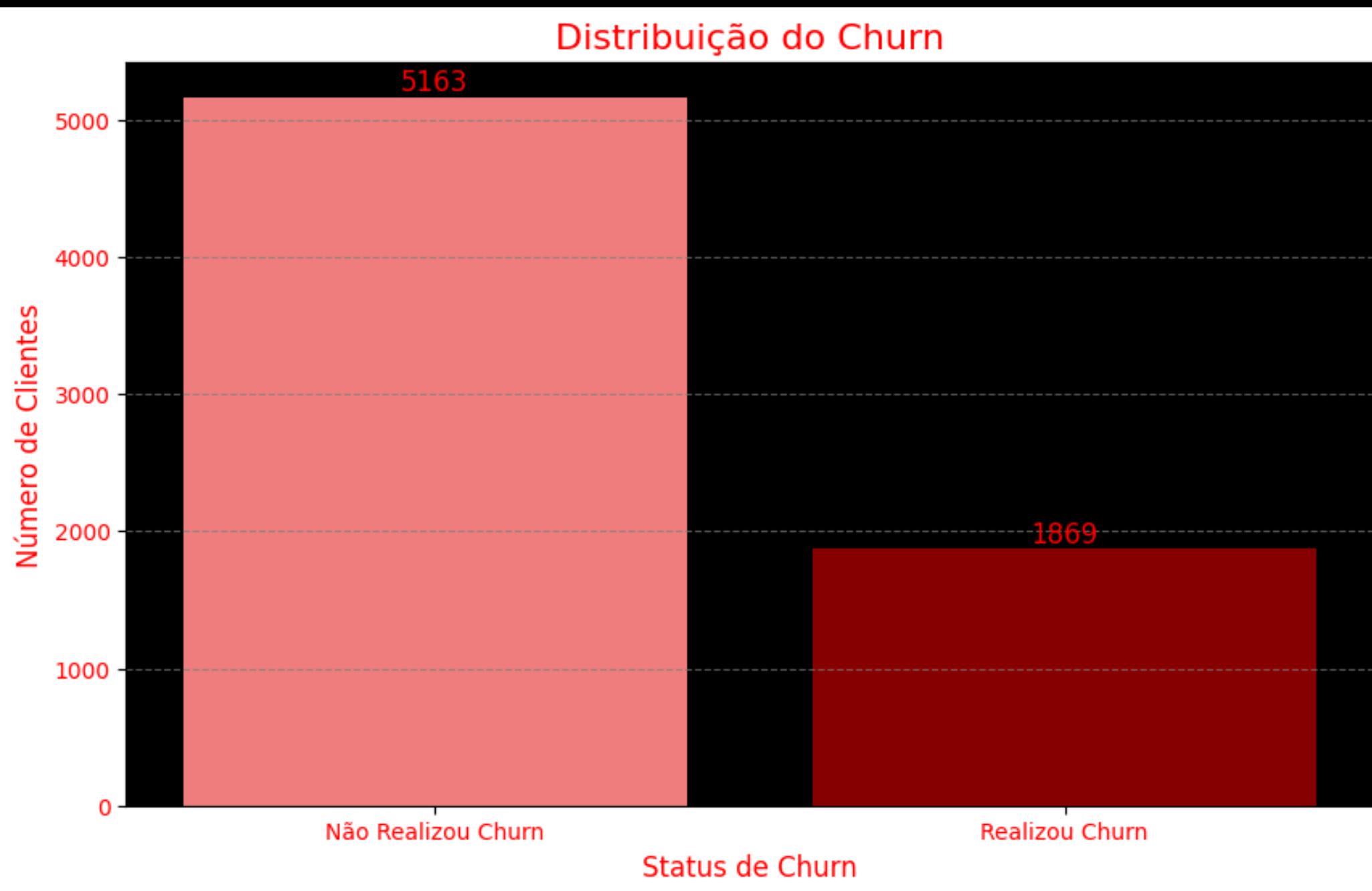
Este gráfico de barras apresenta as 10 variáveis com maior correlação com o churn, classificadas por relevância. Ele ajuda a entender quais fatores têm maior impacto sobre a retenção ou abandono de clientes.

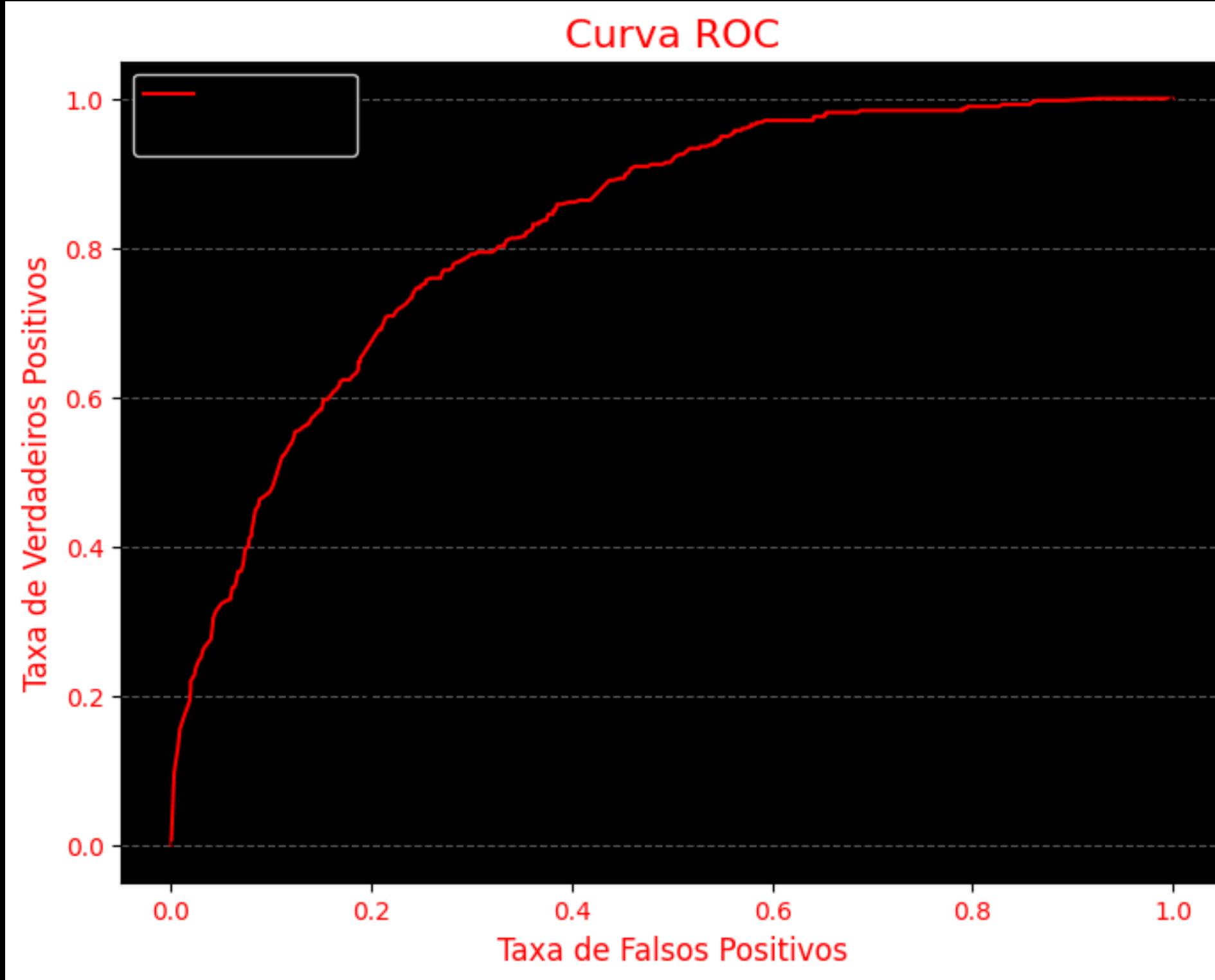
Cada barra representa uma variável, enquanto a altura da barra indica a força da correlação. Por exemplo, variáveis como tipo de contrato ou custo mensal podem ter um impacto significativo no churn.



PRINCIPAIS VARIÁVEIS RELACIONADAS AO CHURN

Este gráfico de barras agrupadas mostra a distribuição de clientes que realizaram ou não o churn. As barras são claramente diferenciadas por cores, com valores visíveis acima de cada barra. O status de churn refere-se à retenção ou perda de clientes. A visualização ajuda a identificar o número de clientes que continuam utilizando o serviço e aqueles que o abandonaram, proporcionando uma visão geral do comportamento dos clientes.

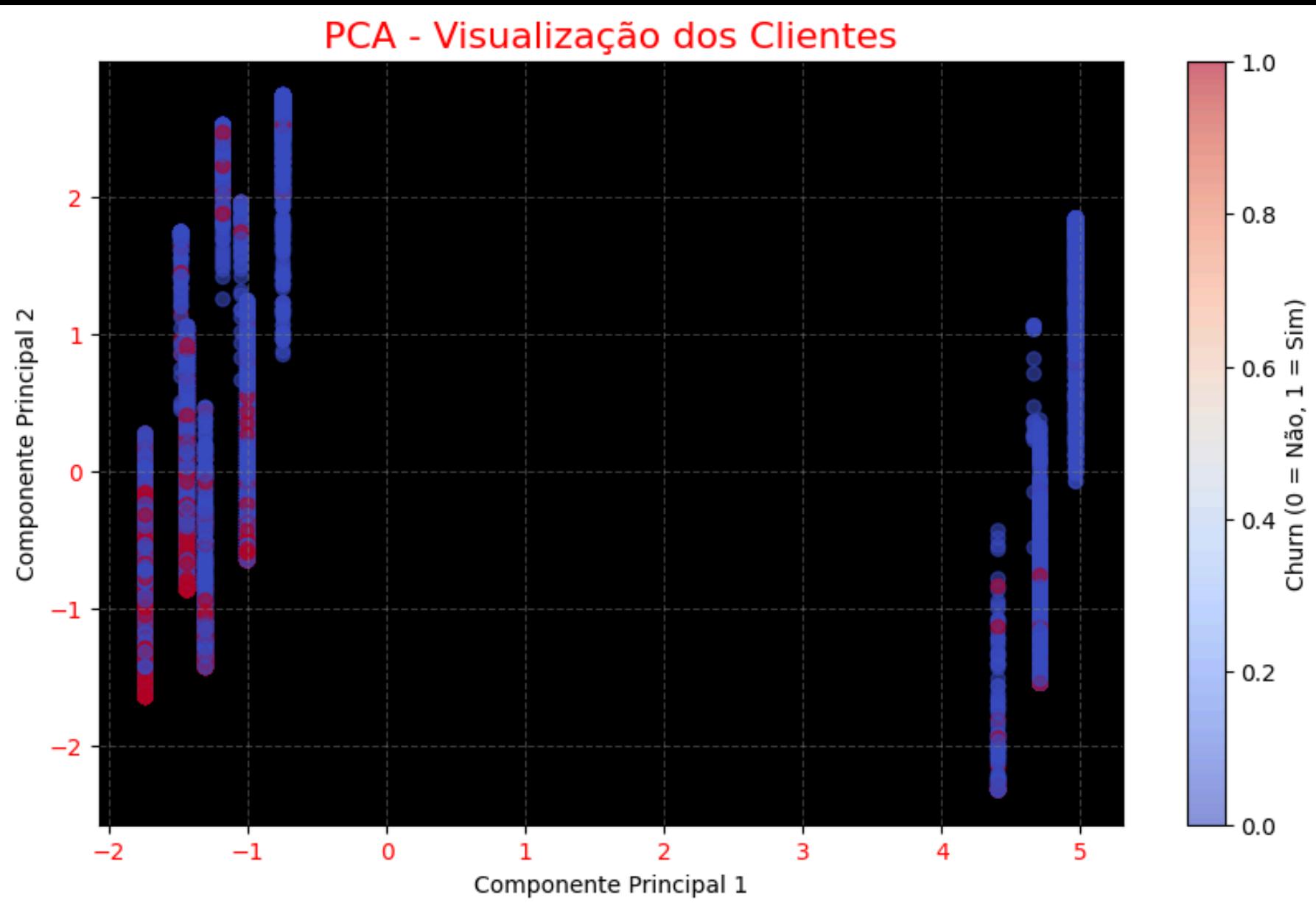




AVALIAÇÃO DE DESEMPENHO DO MODELO: CURVA ROC

A curva ROC mostra o desempenho do modelo de classificação para prever o churn. O eixo X representa a taxa de falsos positivos, enquanto o eixo Y representa a taxa de verdadeiros positivos. Quanto mais próximo da curva está do canto superior esquerdo, melhor é o desempenho do modelo.

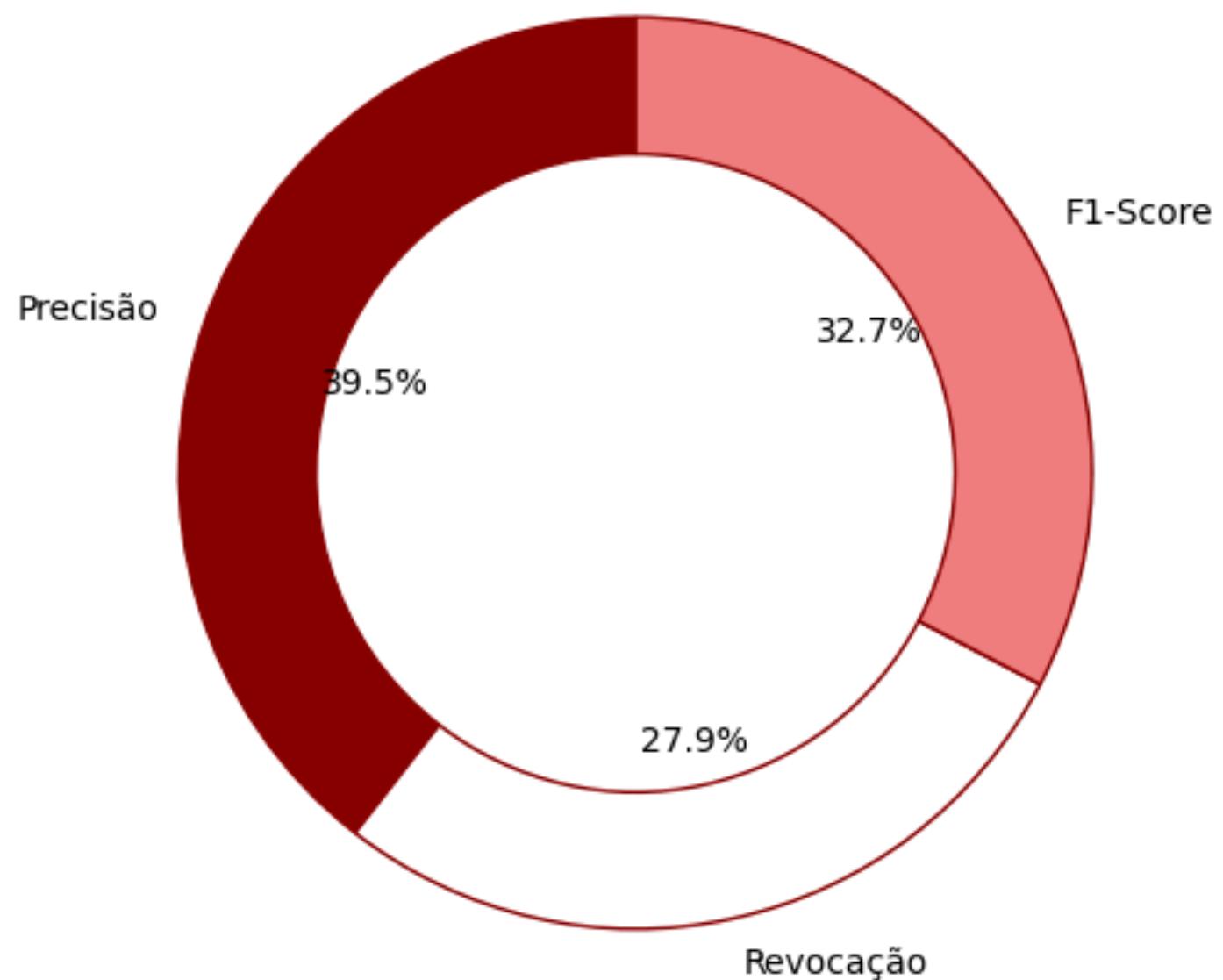
PCA - Visualização dos Clientes



AGRUPAMENTO DE CLIENTES COM E SEM CHURN (PCA)

A Análise de Componentes Principais (PCA) permite reduzir a dimensionalidade dos dados e visualizar os padrões ocultos em duas dimensões. No gráfico acima, cada ponto representa um cliente e as cores indicam se ele realizou churn (vermelho) ou não (azul). É possível observar certa separação entre os grupos, sugerindo que há características distintas entre clientes que permanecem e os que cancelam o serviço. Essa técnica auxilia na compreensão da estrutura dos dados antes da modelagem.

Desempenho do Modelo



ANÁLISE DE DESEMPENHO: PRECISÃO, REVOCAÇÃO E F1-SCORE

Este gráfico de rosca apresenta as métricas principais do modelo de classificação: precisão, revocação e F1-Score. A precisão mede a proporção de previsões corretas entre as classificações positivas. A revocação avalia a capacidade de identificar corretamente os positivos reais. Já o F1-Score é uma média harmônica entre precisão e revocação, útil para balancear falso positivo e falso negativo. Os segmentos destacados por cores proporcionam uma visão clara da contribuição de cada métrica no desempenho geral.



CONCLUSÃO E PRÓXIMOS PASSOS

O modelo é eficaz na predição do churn e pode ser integrado a sistemas de CRM para disparar ações preventivas.

Os próximos passos incluem:

- Validação em dados em tempo real
- Testes com outros algoritmos
- Geração de alertas automatizados para equipes comerciais

REFERÊNCIAS

- IBM Sample Dataset:

<https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/>

- Scikit-Learn: <https://scikit-learn.org/>

- Seaborn Library: <https://seaborn.pydata.org/>

- Matplotlib: <https://matplotlib.org/>



Até a próxima!

OBRIGADO!

2025 CoderHouse - DataScience