1. Introdução

O objetivo deste relatório é documentar a análise e o desenvolvimento de um modelo de Previsão de Churn utilizando um conjunto de dados do setor de telecomunicações. A Previsão de Churn é uma tarefa importante para as empresas, pois permite identificar os clientes que têm maior probabilidade de cancelar seus serviços, permitindo que a empresa tome medidas proativas para retê-los.

---

2. Descrição do Conjunto de Dados

O conjunto de dados utilizado neste projeto é o "Telco Customer Churn". Ele contém informações sobre clientes de uma empresa de telecomunicações, incluindo detalhes sobre os serviços contratados, informações demográficas e se o cliente cancelou ou não o serviço (Churn).

Características do Conjunto de Dados:

- 7043 observações
- 21 variáveis (20 características e 1 variável alvo)
- Tipos de dados: Numérico e Categórico

---

3. Metodologia de Análise

A metodologia de análise avançada nas seguintes etapas:

3.1 Preparação dos Dados:

- Carregar o conjunto de dados
- Dividir os dados em variáveis independentes (X) e variáveis dependentes (y)
- Remover colunas irrelevantes ou duplicadas
- Dividir os dados em conjuntos de treinamento e teste

3.2 Pré-processamento de Dados:

- Tratar valores ausentes: Utilizou-se a mediana para imputar valores ausentes nas variáveis numéricas e a constante 'missing' para variáveis categóricas.
- Codificar variáveis categóricas: Aplicou-se OneHotEncoder para converter variáveis categóricas em variáveis binárias.
- Padronizar variáveis numéricas: Utilizou-se StandardScaler para padronizar as variáveis numéricas.

3.3 Desenvolvimento do Modelo:

- Construir um pipeline de pré-processamento
- Escolher um algoritmo de classificação (RegLog)

- Treinar o modelo nos dados de treinamento
- Avaliar o desempenho do modelo nos dados de teste

3.4 Ajuste de Hiperparâmetros:

- Utilização Grid Search para encontrar os melhores hiperparâmetros

3.5 Avaliação do Modelo:

- Avaliar acurácia, precisão, recall e pontuação F1
- Gerar curva ROC e calcular AUC
- Visualizar a matriz de confusão
- Analisar os resíduos do modelo

---

4. Resultados

4.1 Validação Cruzada:

- A acurácia média da validação cruzada foi de 0.80%.
- E o desvio padrão da acurácia da validação cruzada foi de 0.01%.

4.2 Otimização de Hiperparâmetros:

Utilizou-se GridSearchCV para encontrar os melhores hiperparâmetros para o modelo de regressão logística.

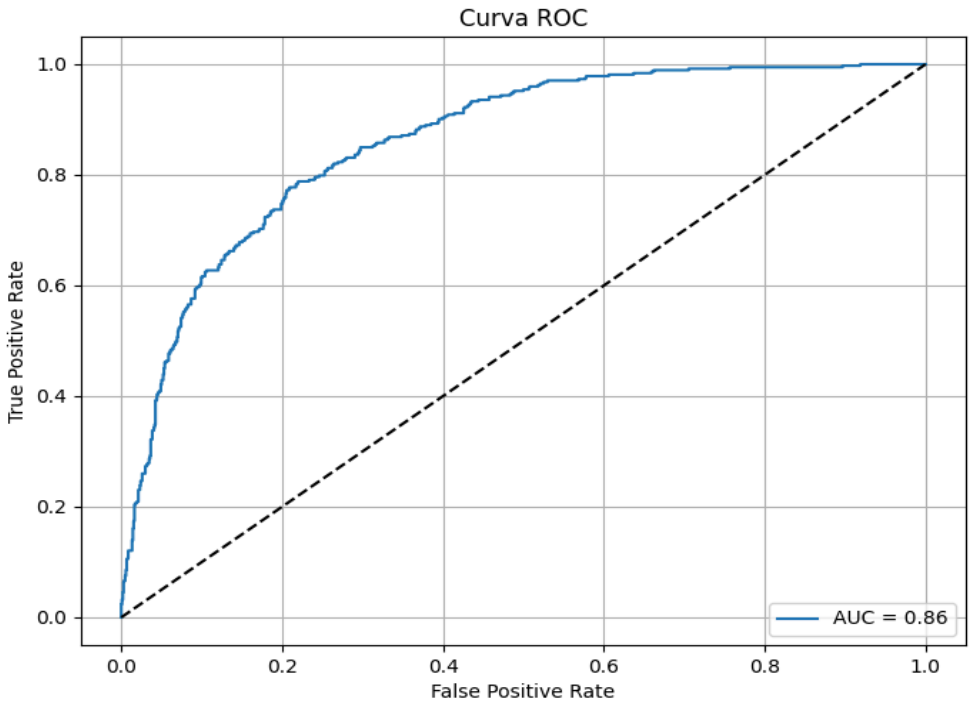
- Melhores Parâmetros Encontrados: {‘classifier_c’: 1, ‘classifier_solver’: ‘liblinear’}

4.3 Desempenho do Modelo:

- O modelo de Churn Prediction obteve uma acurácia de (0.8239886444286728%) nos dados de teste.
- A precisão foi de (0.6947040498442367%), o recall de (0.5978552278820375%) e o F1-score de (0.6426512968299711%).
- Relatório de Classificação:

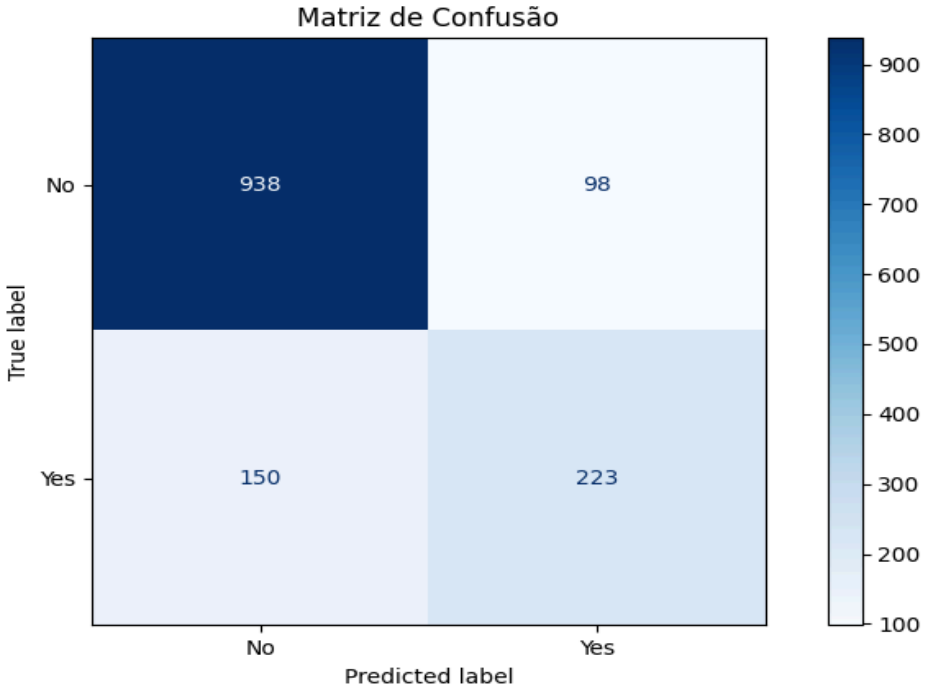
	Precision	**Recall**	**F1-Score**	**Support**
No	0.86	0.91	0.88	1036
Yes	0.69	0.60	0.64	373
Accuracy			0.82	1409
Macro avg	0.78	0.75	0.76	1409
Weighted avg	0.82	0.82	0.82	1409

- A AUC da curva ROC foi de 0.86%



4.4 Matriz de Confusão:

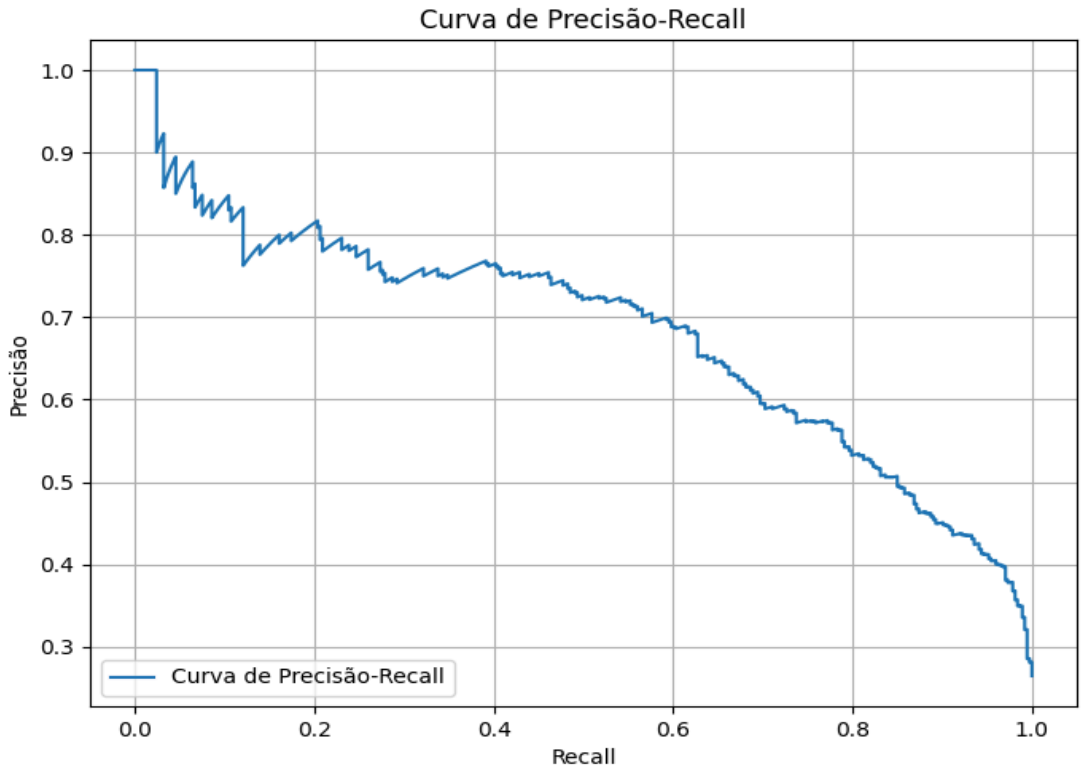
A matriz de confusão fornece uma visão detalhada do desempenho do modelo, destacando as previsões corretas e incorretas em relação às classes reais.



- Um número alto de FNs pode indicar que o modelo precisa ser ajustado para ser mais sensível à classe de churn, possivelmente ajustando os limiares de decisão.
- A matriz de confusão é essencial para entender onde o modelo está errando e tomar medidas corretivas, como coletar mais dados ou ajustar hiperparâmetros.

4.5 Curva de Precisão-Recall

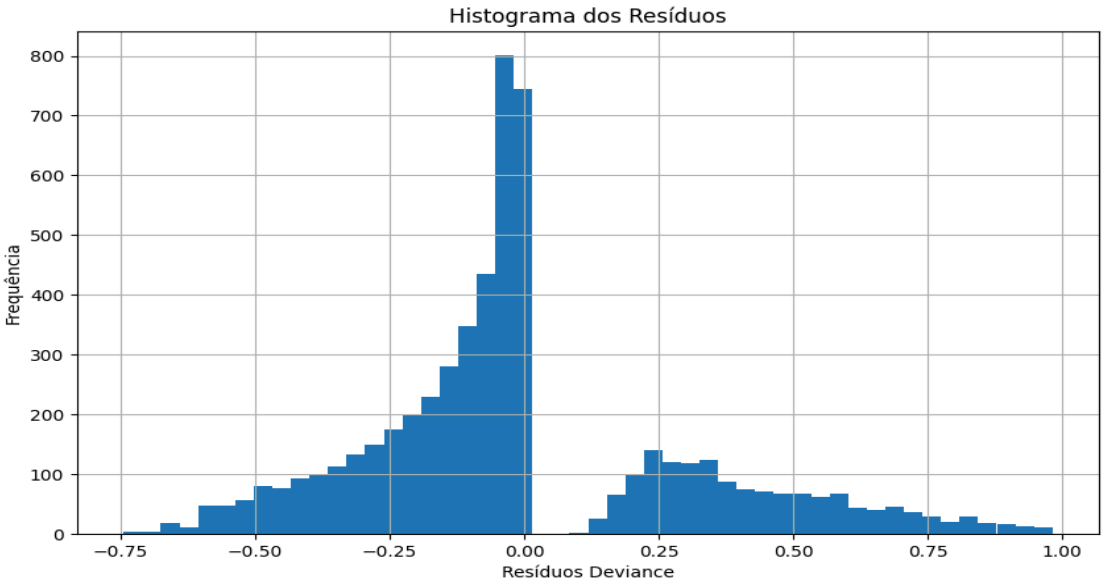
A curva de precisão-recall é especialmente útil para avaliar a performance do modelo em conjuntos de dados desbalanceados, como é comum em problemas de churn, onde a maioria dos clientes pode não cancelar.



- A curva de precisão-recall ajuda a encontrar um equilíbrio entre precisão e recall. Em muitos casos de negócios, é mais importante identificar todos os clientes em risco (alto recall), mesmo que isso resulte em alguns falsos positivos.
- Um modelo ideal teria alta precisão e alto recall, resultando em uma curva que se aproxima do canto superior direito do gráfico.

4.6 Análise de Resíduos

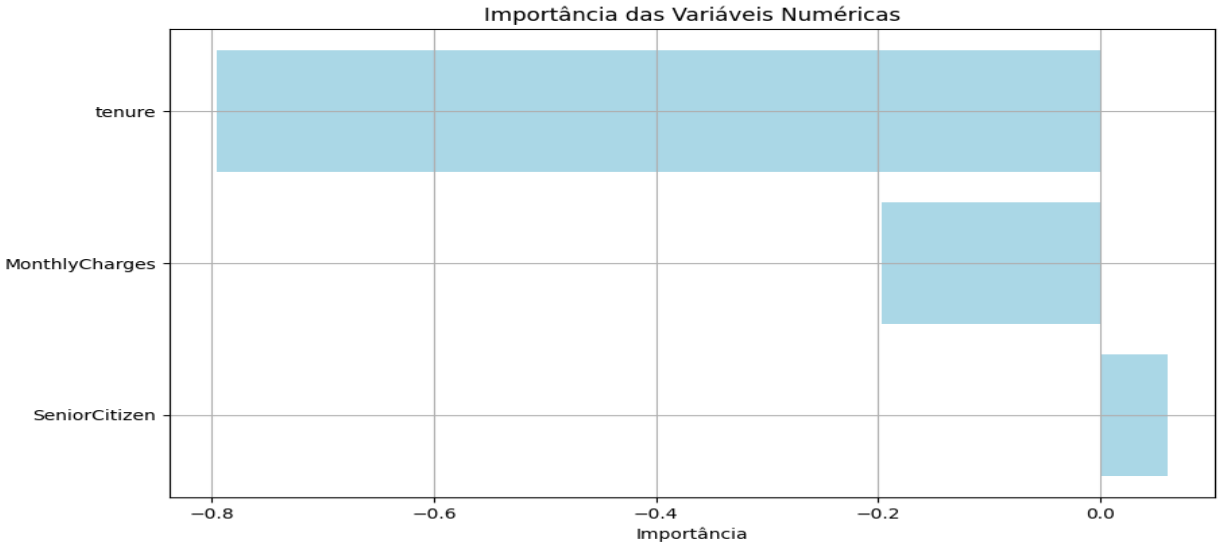
A análise de resíduos examina as diferenças entre os valores previstos pelo modelo e os valores reais observados, ajudando a identificar vieses e problemas no ajuste do modelo.



- Analisar os resíduos ajuda a identificar áreas onde o modelo pode ser melhorado, como coleta de mais dados ou ajustes nos hiperparâmetros.
- Resíduos anômalos podem indicar a necessidade de um modelo mais complexo ou a presença de outliers que precisam ser tratados.

4.7 Importância das Variáveis:

A importância das variáveis é medida pelos coeficientes no modelo de regressão logística. Eles indicam o impacto de cada variável na previsão do modelo.



- Entender as variáveis mais influentes permite direcionar estratégias de retenção. Por exemplo, se o custo é uma variável importante, a empresa pode considerar ajustar sua política de preços ou oferecer descontos.
- A análise da importância das variáveis também pode revelar insights sobre o comportamento do cliente e áreas onde a empresa pode intervir para reduzir o churn.

4.8 Implicações e Aplicações Potenciais:

- O modelo desenvolvido pode ser usado para identificar clientes com maior risco de churn
- As informações obtidas podem ser usadas para desenvolver estratégias de retenção de clientes e melhorar a satisfação do cliente
- A empresa pode economizar recursos direcionando suas campanhas de marketing e ofertas promocionais para os clientes certos

---

5. Conclusão

Neste projeto, foi desenvolvido um modelo de Previsão de Churn usando um conjunto de dados do setor de telecomunicações. O modelo apresentou um desempenho superior na previsão de rotatividade de clientes, e suas descobertas têm implicações importantes para a empresa, fornecendo insights importantes para ações futuras de retenção de clientes.