PROJETO 1 - Análise de Churn:

Segmentação (com K-Means) e Predição (com Regressão Logística)

📊 Introdução

Este projeto tem como objetivo analisar a evasão de clientes (churn) a partir de um dataset processado na plataforma Knime. O projeto busca aplicar técnicas de **clusterização** e **regressão logística** para compreender melhor o comportamento dos clientes e prever aqueles com maior propensão ao cancelamento. A clusterização permitirá segmentar a base de clientes com base em padrões de uso, perfil demográfico e histórico de pagamentos, ajudando a identificar grupos mais suscetíveis ao churn. Em seguida, a regressão logística será utilizada para quantificar o impacto de diferentes variáveis na decisão de cancelamento, permitindo a construção de um modelo preditivo. Com esses insights, a empresa poderá direcionar ações mais eficazes para retenção, personalizando ofertas e otimizando a experiência dos clientes de forma estratégica.

💡 Problema de Negócio

A taxa de churn de 15% entre os clientes da empresa de telecomunicações iraniana representa um desafio significativo para a sustentabilidade do negócio. Esse nível de cancelamento pode indicar insatisfação com os serviços, concorrência acirrada ou inadequação dos planos oferecidos às necessidades dos clientes.

Além do impacto direto na receita, a perda de clientes aumenta os custos de aquisição para reposição e pode afetar a reputação da empresa no mercado. Para mitigar esse problema, é essencial identificar os fatores que influenciam o churn e desenvolver estratégias preditivas e preventivas para retenção, como personalização de ofertas, melhorias no atendimento e programas de fidelização.

🛠️ Solução

Para isso, foi aplicada a técnica de *K-means* para segmentar os clientes em diferentes clusters, permitindo uma análise mais granular dos padrões de comportamento. Em seguida, foi utilizada a regressão logística em cada segmento para prever a probabilidade de churn com base em variáveis como tempo de assinatura, frequência de uso e valor do cliente.

A análise busca identificar os principais fatores que influenciam a saída dos clientes, tanto no conjunto geral quanto em cada cluster específico. A interpretação dos coeficientes do modelo permite entender o impacto de cada variável na decisão de churn. Além disso, métricas de desempenho, como matriz de confusão, precisão e recall, foram avaliadas para medir a eficácia dos modelos. Com esses insights, espera-se fornecer recomendações estratégicas para reduzir a evasão e melhorar a retenção de clientes.

📂 Fonte de Dados

O dataset pode ser acessado pelo link: https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset

Interpretação dos Clusters

A seguir são listadas as principais observações sobre os resultados da clusterização:

Cluster 0 (clientes com valores positivos na maioria das variáveis):

- **Alta frequência de uso** (+0.88) e **maior tempo de uso** (+0.87)

- **Mais ligações distintas feitas** (+0.65)
- **Maior duração da assinatura** (+0.16)
- **Pagam valores mais altos** (+0.62)
- **Mais propensos a reclamar** (-0.21, ou seja, abaixo da média, mas próximo de 0)
- **Idade e valor do cliente acima da média**

🟢 **Possível perfil:**
Este cluster pode representar **clientes fiéis e de alto valor**, que usam bastante os serviços e têm contratos longos. Como reclamam pouco, podem ter uma boa experiência com a empresa. O churn aqui pode ser baixo.

Cluster 1 (clientes com valores negativos na maioria das variáveis):

- **Menor frequência de uso** (-0.54) e **menos tempo de uso** (-0.53)
- **Fazem menos ligações distintas** (-0.40)
- **Assinaturas mais curtas** (-0.10)
- **Pagam valores mais baixos** (-0.38)
- **Mais propensos a reclamar** (+0.13)
- **Idade abaixo da média**

🔴 **Possível perfil:**
Este cluster pode representar **clientes de baixo engajamento e mais propensos ao churn**. Eles usam menos os serviços, gastam menos e podem estar insatisfeitos.

Interpretação da Regressão Logística

A seguir um quadro comparativo com a interpretação detalhada das variáveis para todos os segmentos (geral e clusters_0 e cluster_1).

Comparação da Regressão Logística por Segmento

Variável	Geral	Cluster 0	Cluster 1
Call Failure	Coeficiente negativo, mas não significativo (p>0,05). Pequena relação com churn.	Coeficiente negativo, mas também não significativo. Nenhuma evidência de impacto sobre o churn.	Coeficiente negativo e altamente significativo (p=0). Falhas de chamadas aumentam muito a probabilidade de churn.

Complains	Coeficiente muito negativo e significativo (p≈0). Reclamações aumentam drasticamente o churn.	Mesmo comportamento da análise geral: impacto fortemente negativo e significativo. Leads que reclamam tendem a sair.	Coeficiente negativo e altamente significativo (p=0). Reclamações são um indicador crítico de churn.
Subscription Length	Coeficiente negativo e significativo (p<0,05). Clientes com assinaturas mais longas tendem a permanecer.	Impacto ainda mais forte do que no geral. Assinaturas longas reduzem o churn de forma relevante.	Não significativo (p=0,79). Neste cluster, tempo de assinatura não influencia a retenção.
Charge Amount	Coeficiente positivo e significativo (p<0,01). Clientes que gastam mais são menos propensos ao churn.	Mesmo padrão, mas com efeito mais forte. Gastos elevados estão associados à retenção.	Coeficiente positivo e altamente significativo. Aumento nos gastos reduz a probabilidade de churn.
Seconds of Use	Coeficiente positivo e significativo. Quanto maior o uso, menor a taxa de churn.	Coeficiente ainda mais forte que no geral. Uso intensivo reduz significativamente a chance de churn.	Coeficiente negativo e significativo. Comportamento oposto: aumento no uso pode indicar insatisfação.
Frequency of Use	Coeficiente positivo e significativo. Uso frequente está relacionado à maior retenção.	Mesma relação encontrada na análise geral. Usuários frequentes tendem a permanecer.	Impacto fortemente positivo e significativo. Frequência de uso reduz o churn.
Frequency of SMS	Coeficiente positivo e significativo. Envio frequente de SMS reduz o churn.	Impacto positivo e significativo. Clientes que enviam mais SMS tendem a ficar.	Não significativo (p=0,84). No Cluster 2, a frequência de SMS não influencia o churn.
Distinct Called Numbers	Coeficiente positivo e significativo. Contatar mais números reduz a chance de churn.	Impacto fortemente positivo e altamente significativo.	Não significativo (p=0,14). Neste cluster, diversidade de contatos não tem relação com churn.
Age Group	Coeficiente positivo e significativo. Grupos etários mais velhos têm menor churn.	Impacto ainda mais forte do que na análise geral. Idade é um fator relevante para retenção.	Não significativo (p=0,90). Neste cluster, a idade não influencia a saída.

Tariff Plan	Coeficiente positivo e significativo. Determinados planos tarifários aumentam a retenção.	Mesmo comportamento da análise geral.	Coeficiente negativo e marginalmente significativo (p≈0,05). Alguns planos podem estar associados ao churn.
Age	Coeficiente negativo e altamente significativo. Clientes mais velhos têm menor churn.	Mesmo padrão, mas com efeito ainda mais forte do que na análise geral.	Não significativo (p=0,31). Idade isolada não influencia este cluster.
Customer Value	Coeficiente negativo e significativo. Clientes com menor valor de vida têm maior churn.	Mesmo comportamento da análise geral. Clientes de baixo valor são mais propensos a sair.	Coeficiente positivo e marginalmente significativo (p=0,08). Clientes de alto valor podem ter maior retenção.
Intercept (Constante)	Coeficiente positivo e significativo. Grupo de referência tem baixa propensão ao churn.	Mesmo comportamento da análise geral.	Mesmo comportamento da análise geral.

Principais Conclusões - Análise Geral e por Segmento

📌 **Análise Geral (Sem Segmentação):**

- Clientes que reclamam (**Complains**) têm a maior relação com churn, sendo um fator altamente significativo.
- Gastos maiores (**Charge Amount**) e tempo de assinatura (**Subscription Length**) reduzem a chance de churn, pois refletem maior envolvimento com a empresa.
- Uso de serviços (**Seconds of Use** e **Frequency of Use**) é um preditor forte de retenção – clientes mais ativos tendem a permanecer.
- Clientes mais velhos e com maior **Customer Value** apresentam menor churn.
- Planos tarifários afetam o churn, sugerindo que alguns planos podem ser mais atrativos para retenção.

Cluster 0:

- Padrões semelhantes à análise geral, mas com efeitos ainda mais intensos em algumas variáveis.
- Leads que assinam por mais tempo, gastam mais e fazem mais ligações têm probabilidade muito menor de churn.

- Idade e diversidade de contatos são fatores importantes para retenção, indicando que leads mais sociáveis e experientes tendem a permanecer.
- Foco na retenção: evitar churn por reclamações e baixo gasto.

Cluster 1:

- Diferente da análise geral, uso intenso de minutos está ligado a maior churn, o que pode indicar insatisfação.
- **Call Failures** e **Complaints** são os maiores preditores de churn (efeitos muito mais fortes que nos outros segmentos).
- Planos tarifários podem estar associados ao churn, sugerindo que algumas ofertas podem não estar alinhadas às expectativas desse grupo.

Diferente dos outros segmentos, **frequência de SMS** e **idade** não afetam a retenção.

Conclusões Estratégicas

✅ **Para a base como um todo (sem segmentação):**

- Minimizar reclamações deve ser a prioridade número 1 para reduzir churn.
- Incentivar maior gasto e engajamento melhora a retenção, pois clientes que investem mais tendem a permanecer.
- Oferecer planos personalizados para clientes mais antigos e de alto valor, já que esses fatores reduzem o churn.

✅ **Para Cluster 1:**

- Foco em clientes com alta retenção natural → reforçar benefícios para quem gasta mais e tem longa assinatura.
- Melhorar atendimento ao cliente para evitar churn por reclamações.
- Explorar planos diferenciados para clientes mais experientes, pois idade e diversidade de contatos influenciam a retenção.

✅ **Para Cluster 2:**

- Identificar e resolver problemas de serviço (**Call Failures** e **Complaints**) rapidamente.
- Analisar o impacto de planos tarifários – pode ser necessário reformular ofertas para esse grupo.
- Avaliar o motivo pelo qual uso intenso de minutos está correlacionado ao churn, pois isso não ocorre nos outros segmentos.

Considerações Finais

Os resultados deste estudo forneceram insights valiosos sobre os fatores que influenciam o churn, tanto de forma geral quanto segmentados por clusters. A análise revelou que variáveis como tempo de assinatura, valor gasto e frequência de uso têm forte impacto na retenção de clientes.

A segmentação por K-Means permitiu identificar padrões distintos de comportamento, evidenciando que diferentes perfis de clientes possuem motivações variadas para cancelar o serviço. A regressão logística aplicada a cada cluster mostrou que os fatores determinantes do churn variam entre os grupos, reforçando a necessidade de estratégias personalizadas.

Com base nesses achados, recomenda-se que a empresa adote ações direcionadas a cada perfil de cliente, aumentando a retenção e reduzindo o churn. A análise também sugere a importância de aprimorar o atendimento e ajustar planos tarifários conforme o comportamento dos usuários. Por fim, futuras análises podem explorar modelos mais avançados para prever churn com maior precisão.

Como analisar os resultados da regressão logística

1. **Variable**
    - Nome da variável independente usada na regressão logística. Essas são as features que impactam a variável dependente.
2. **Coeff.**
    - Coeficiente da regressão logística.
    - **Interpretação**: Indica o impacto da variável independente sobre a variável dependente. Se for **positivo**, a variável aumenta a probabilidade do evento ocorrer (churn, por exemplo). Se for **negativo**, reduz essa probabilidade.
3. **Std. Err.**
    - Erro padrão do coeficiente estimado.
    - **Interpretação**: Mede a incerteza na estimativa do coeficiente. Quanto menor, mais confiável é a estimativa.
4. **z-score**
    - Estatística de teste para avaliar a significância da variável.
    - **Interpretação**: Calculado como Coeff.Std. Err.\frac{\text{Coeff.}}{\text{Std. Err.}}Std. Err.Coeff.. Quanto mais alto (positivo ou negativo), mais significativa é a variável para o modelo.
5. **P>|z|**
    - Valor-p para o teste de significância da variável.
    - **Interpretação**: Se **P>|z|** for pequeno (geralmente menor que 0.05), significa que a variável tem um impacto estatisticamente significativo no modelo.

Como analisar esses elementos?

1. **Identificar variáveis significativas**: Verifique o **P>|z|**. Se for menor que 0.05, a variável é estatisticamente significativa para prever o churn.
2. **Interpretar os coeficientes**: Coeficientes positivos aumentam a chance de churn, enquanto negativos diminuem.
3. **Verificar a estabilidade dos coeficientes**: Valores altos de **Std. Err.** indicam instabilidade na estimativa.
4. **Avaliar a força do efeito**: O **z-score** ajuda a entender quais variáveis têm maior impacto na previsão.

Esses elementos são métricas comuns para avaliar o desempenho de um modelo de classificação, incluindo a **regressão logística** no seu projeto no **KNIME**. Aqui está a descrição de cada um deles:

Métricas de Matriz de Confusão

1. **True Positives (TP)**
   - Quantidade de previsões corretas onde o modelo **previu churn** e o cliente realmente cancelou.
   - **Exemplo**: Se 50 clientes cancelaram e o modelo previu corretamente o churn para 30 deles, então **TP = 30**.
2. **False Positives (FP)**
   - Número de vezes que o modelo **previu churn** erroneamente para clientes que **não cancelaram**.
   - **Exemplo**: Se 20 clientes foram marcados como churn pelo modelo, mas permaneceram na empresa, então **FP = 20**.
   - Também chamado de **erro tipo I**.
3. **True Negatives (TN)**
   - Número de previsões corretas onde o modelo **previu que o cliente não cancelaria** e ele realmente não cancelou.
   - **Exemplo**: Se 500 clientes continuaram com a empresa e o modelo previu corretamente para 450 deles, então **TN = 450**.
4. **False Negatives (FN)**
   - Número de vezes que o modelo **não previu churn**, mas o cliente cancelou.
   - **Exemplo**: Se 50 clientes cancelaram, mas o modelo só identificou 30, então os outros 20 foram classificados erroneamente como "não churn", logo **FN = 20**.
   - Também chamado de **erro tipo II**.

Métricas Derivadas da Matriz de Confusão

5. **Recall (Sensibilidade ou Taxa de Verdadeiros Positivos)**
    - Mede a capacidade do modelo de capturar corretamente os clientes que realmente cancelaram.
    - **Fórmula**: Recall=TPTP+FNRecall = \frac{TP}{TP + FN}Recall=TP+FNTP
    - **Interpretação**: Quanto maior o Recall, melhor o modelo está identificando os clientes que realmente cancelam.
6. **Precision (Precisão)**
    - Mede a qualidade das previsões positivas do modelo.
    - **Fórmula**: Precision=TPTP+FPPrecision = \frac{TP}{TP + FP}Precision=TP+FPTP
    - **Interpretação**: Indica quantas das previsões de churn realmente correspondem a clientes que cancelaram.
7. **Sensitivity (Sensibilidade)**
    - **Mesmo que o Recall**, ou seja, a capacidade do modelo de identificar corretamente os churns.
8. **Specificity (Especificidade)**
    - Mede a capacidade do modelo de prever corretamente os clientes que **não cancelam**.
    - **Fórmula**: Specificity=TNTN+FPSpecificity = \frac{TN}{TN + FP}Specificity=TN+FPTN
    - **Interpretação**: Se a Specificity for alta, o modelo evita falsos alarmes (prever churn quando o cliente na verdade não cancela).
9. **F-measure (F1-score)**
    - Média harmônica entre **Recall** e **Precision**, usada quando há um **desequilíbrio entre classes** (muito mais clientes que não cancelam do que clientes que cancelam).
    - **Fórmula**: F1=2×Precision×RecallPrecision+RecallF1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}F1=2×Precision+RecallPrecision×Recall
    - **Interpretação**: Um valor alto indica um bom equilíbrio entre precisão e sensibilidade.
10. **Accuracy (Acurácia)**

- Mede a proporção total de previsões corretas.
- **Fórmula**: Accuracy=TP+TNTP+TN+FP+FNAccuracy = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN
- **Interpretação**: Pode ser enganosa se o dataset for desbalanceado (exemplo: se 90% dos clientes não cancelam, um modelo que sempre prevê "não churn" terá 90% de acurácia, mas será inútil).

11. **Cohen's Kappa**

- Mede o nível de concordância entre as previsões do modelo e os valores reais, levando em conta a concordância esperada ao acaso.

- **Fórmula**: Kappa=Po−Pe1−PeKappa = \frac{P_o - P_e}{1 - P_e}Kappa=1−PePo−Pe Onde:
    - PoP_oPo = Proporção de concordância observada
    - PeP_ePe = Concordância esperada pelo acaso
- **Interpretação**:
    - **Kappa = 1** → Concordância perfeita
    - **Kappa > 0.8** → Muito bom
    - **Kappa entre 0.6 e 0.8** → Bom
    - **Kappa entre 0.4 e 0.6** → Moderado
    - **Kappa < 0.4** → Fraco

Resumo para Análise

1. Se **Recall for alto**, o modelo identifica bem quem cancela, mas pode ter falsos positivos.
2. Se **Precision for alta**, as previsões de churn são mais confiáveis, mas pode perder alguns clientes que realmente cancelam.
3. Se **F1-score for alto**, há um bom equilíbrio entre Recall e Precision.
4. Se **Accuracy for alta**, o modelo acerta no geral, mas pode ser enganador se os dados forem desbalanceados.
5. Se **Cohen's Kappa for alto**, significa que o modelo está realmente agregando valor acima do puro acaso.