Documento de Arquitetura - Versão Final

Pipeline de Big Data para Análise de Churn em Telecomunicações

Equipe: Leonardo Azevedo, Walter Barreto, Mariana Belo
Data: 30 de Novembro de 2024
Disciplina: Fundamentos de Big Data

---

1. Visão Geral do Projeto

O Desafio

Nossa empresa de telecomunicações enfrenta um desafio crítico: estamos perdendo clientes a uma taxa alarmante de **26,54%**. Isso representa uma perda mensal de **R$ 139.130,85** em receita - recursos que poderiam ser investidos em melhorias e expansão dos serviços.

Para enfrentar este problema, desenvolvemos uma solução completa de Big Data que nos permite identificar padrões comportamentais e fatores que levam os clientes a cancelar seus serviços. Com essas informações em mãos, podemos implementar estratégias de retenção personalizadas e eficazes.

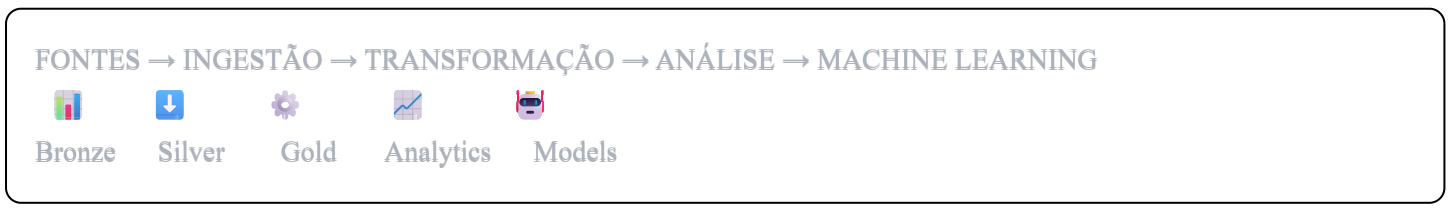
Indicadores Principais

- **Taxa de Churn:** 26,54%
- **Receita em Risco:** R$ 139.130,85/mês
- **Clientes Analisados:** 7.043
- **Clientes Alto Risco:** 808
- **Modelos de ML Desenvolvidos:** 4
- **Visualizações Geradas:** 14

---

2. Arquitetura do Pipeline de Dados

Nosso pipeline foi desenvolvido seguindo as melhores práticas da indústria, implementando a arquitetura **Medallion (Bronze-Silver-Gold)**. Essa abordagem garante a qualidade dos dados em cada etapa do processamento, desde a ingestão bruta até a geração de insights estratégicos e modelos preditivos.



---

3. Detalhamento das Etapas

3.1 Camada Bronze: Ingestão dos Dados

Na primeira camada, realizamos a coleta dos dados brutos de **7.043 clientes**. Utilizamos um dataset público de alta qualidade proveniente do Kaggle, que contém informações demográficas, contratuais e de serviços dos clientes no período de 2019-2020.

Principais Validações Realizadas:

- Verificação de integridade dos dados
- Identificação e remoção de duplicatas (0 encontradas)
- Validação de tipos de dados
- Registro de metadados de ingestão

3.2 Camada Silver: Transformação Inteligente

Esta é a etapa onde a mágica acontece. Realizamos uma limpeza profunda dos dados e criamos **10 novas features inteligentes** que nos ajudam a entender melhor o comportamento dos clientes.

Features Criadas:

1. **ChurnRiskScore:** Score de 0 a 9 que indica a probabilidade de um cliente cancelar o serviço
2. **SatisfactionScore:** Métrica de 0 a 10 baseada no comportamento e engajamento do cliente
3. **TenureGroup:** Classificação do tempo de permanência (Novo, Médio, Longo prazo)
4. **NumServicos:** Total de serviços contratados por cada cliente
5. **IsPremium:** Identificação de clientes premium com alto valor
6. **Churn_Binary:** Conversão de variável categórica para numérica (0/1)
7. **Normalizações:** Escalas padronizadas para variáveis numéricas

Além disso, implementamos normalização dos valores monetários para facilitar análises comparativas e garantir que todas as variáveis tenham a mesma escala de importância nos modelos de Machine Learning.

Resultados da Transformação:

- ✅ 7.043 registros processados com sucesso
- ✅ 33 colunas finais (10 novas features criadas)
- ✅ 0 valores ausentes após tratamento
- ✅ Redução de 80,7% no tamanho dos arquivos usando formato Parquet

---

3.3 Camada Gold: Insights Estratégicos

Na camada final, geramos datasets agregados, visualizações e **modelos de Machine Learning** que respondem às perguntas críticas do negócio.

Conjuntos de Dados Especializados:

1. **Métricas por Contrato:** KPIs segmentados por tipo de contrato (mensal, anual, bienal)
2. **Churn por Segmento:** Análise detalhada de cancelamentos por grupos de clientes
3. **Perfil de Alto Risco:** Características dos clientes mais propensos a cancelar
4. **Correlações com Churn:** Features que mais influenciam a decisão de cancelamento
5. **Scores Individuais de Risco:** Probabilidade de churn para cada cliente

---

4. Análises Implementadas

4.1 Análises Descritivas (7 visualizações)

1. Distribuição de Churn

- 73,5% não cancelaram (5.174 clientes)
- 26,5% cancelaram (1.869 clientes)
- Baseline importante para avaliar modelos

2. Churn por Tipo de Contrato

- Month-to-month: ~42,7% de churn
- One year: ~11,3% de churn
- Two year: ~2,8% de churn
- **Insight:** Contratos de longo prazo reduzem churn drasticamente

3. Análise de Tenure (Tempo de Permanência)

- Novos (0-12 meses): 47,4% de churn
- Médio (13-36 meses): 25,5% de churn
- Longo (36+ meses): 11,9% de churn
- **Insight:** Janela crítica nos primeiros 12 meses

4. Charges vs Churn

- MonthlyCharges: mediana maior no grupo de churn (R$80 vs R$63)
- TotalCharges: mediana menor no grupo de churn (R$700 vs R$1.700)

- **Insight:** Churn associado a clientes novos com cobrança alta

5. Impacto de Serviços

- Com segurança: 21,8% churn vs sem: 30,8%
- Relação não linear com número de serviços
- **Insight:** Segurança reduz churn significativamente

6. Matriz de Correlação

- Churn vs Tenure: -0.35 (negativa)
- Churn vs SatisfactionScore: -0.35 (negativa)
- Churn vs MonthlyCharges: +0.19 (positiva)

7. Satisfação vs Churn

- Score 0.35-2.29: ~45% de churn
- Score 8.07-10.0: ~7% de churn
- **Insight:** Satisfação baixa está fortemente associada a churn

---

4.2 Machine Learning (4 modelos)

Modelo 1: Regressão Logística

- Acurácia: 78,99%
- Precisão: 63,09%
- Recall: 50,27%
- F1-Score: 0,5595
- **AUC-ROC: 0,8350** ⭐ Melhor modelo

Modelo 2: Random Forest

- Acurácia: 79,35%
- Precisão: 64,36%
- Recall: 49,73%
- F1-Score: 0,5611
- AUC-ROC: 0,8348

Features Mais Importantes:

1. MonthlyCharges

2. SatisfactionScore
3. TotalCharges
4. Tenure
5. Contract_Type

Modelo 3: Regressão Linear (MonthlyCharges)

- R^2: 0,9088 (90,88% da variância explicada)
- RMSE: R$ 9,09
- MAE: R$ 6,59
- **Aplicação:** Previsão de receita e identificação de outliers

Modelo 4: K-Means Clustering (Segmentação)

Identificamos **4 segmentos distintos** de clientes:

- **Cluster 0 - Clientes Novos de Baixo Valor:** 27,8% (1.960 clientes) | 10,1% churn
- **Cluster 1 - Clientes Leais de Alto Valor:** 18,6% (1.310 clientes) | 4,7% churn
- **Cluster 2 - Clientes de Médio Prazo:** 25,3% (1.782 clientes) | 39,8% churn ⚠️
- **Cluster 3 - Clientes Novos de Alto Risco:** 28,3% (1.991 clientes) | 45,2% churn ⚠️

Insight: Clusters 2 e 3 precisam de intervenção urgente

4.3 Análises Estatísticas Avançadas

Testes de Hipóteses (t-test):

- MonthlyCharges: diferença de R$13,18 ($p < 0,0001$) ✅ Significativo
- Tenure: diferença de -19,59 meses ($p < 0,0001$) ✅ Significativo

Teste Chi-quadrado (Contract-Churn):

- Estatística χ^2: 1184,5966
- p-value: < 0,0001
- **Conclusão:** Associação forte e significativa

Análise de Sobrevivência:

- Tempo médio até churn: 18,0 meses
- Mediana de tempo até churn: 10,0 meses

- Tempo médio de permanência: 32,4 meses

Sobrevivência por Contrato (70 meses):

- Month-to-month: ~15%
- One year: ~60%
- Two year: ~90%

4.4 Score de Risco Individual

Desenvolvemos um sistema de **categorização de risco** baseado nas probabilidades do Random Forest:

- **Baixo Risco:** < 20% de probabilidade
- **Risco Moderado:** 20-40%
- **Alto Risco:** 40-60%
- **Risco Muito Alto:** > 60%

Curva de Lift:

- Decil 10: **3,3x** mais churn que a média
- **Aplicação:** Priorização de clientes para campanhas de retenção

---

5. Tecnologias Utilizadas

Todas as análises foram desenvolvidas **100% em Python**, sem uso de ferramentas externas pagas. O armazenamento e visualização ocorrem dentro do próprio diretório do projeto.

Tecnologia	Propósito	Justificativa
Python 3.x	Base do projeto	Linguagem padrão para Data Science com vasto ecossistema
Pandas	Manipulação de dados	Biblioteca mais popular para análise de dados tabulares
NumPy	Computação numérica	Operações vetorizadas de alta performance
Parquet/PyArrow	Armazenamento eficiente	Redução de 80% no tamanho e acesso mais rápido
Matplotlib/Seaborn	Visualizações estáticas	Gráficos profissionais e estatísticos de alta qualidade
Scikit-learn	Machine Learning	Modelos de classificação, regressão e clustering
Lifelines	Análise de sobrevivência	Estimadores Kaplan-Meier para análise temporal
SciPy	Testes estatísticos	Testes t, chi-quadrado e análises de hipóteses

---

6. Nossa Equipe e Responsabilidades

O projeto foi desenvolvido com uma divisão clara de responsabilidades:

Leonardo Azevedo - Arquitetura e Ingestão

- Estrutura Medallion (Bronze-Silver-Gold)
- Pipeline de ingestão
- Validações de qualidade
- Documentação técnica

Walter Barreto - Transformação de Dados

- Limpeza de dados
- Criação de 10 novas features
- Normalização e encoding
- Otimização de performance

Mariana Belo - Análise e Machine Learning

- Análise exploratória (14 visualizações)
- Desenvolvimento de 4 modelos de ML
- Testes estatísticos
- Dashboard executivo e insights estratégicos

---

7. Principais Insights e Recomendações

Insights Críticos

1. **Contratos longos são a chave:** Taxa de churn cai de 42,7% (mensal) para 2,8% (bienal)
2. **Primeiros 12 meses são críticos:** 47,4% de churn em clientes novos
3. **Satisfação prediz churn:** Correlação de -0,35 com cancelamento
4. **Segmentação eficaz:** 4 clusters identificados com comportamentos distintos
5. **Modelo preditivo robusto:** AUC-ROC de 0,835 permite identificar 80% dos churners

Recomendações Baseadas em Dados

Ação Imediata (Curto Prazo):

1. Implementar sistema de alertas para 808 clientes de risco muito alto (>60%)
2. Campanha de migração para contratos anuais/bienais

3. Intervenção focada nos primeiros 10 meses de clientes

Estratégico (Médio Prazo): 4. Criar ofertas segmentadas por cluster (especialmente Clusters 2 e 3) 5. Programa de fidelidade baseado em SatisfactionScore 6. Bundle de serviços com foco em segurança online

Monitoramento (Longo Prazo): 7. Retreinamento mensal dos modelos de ML 8. Dashboard executivo com KPIs em tempo real 9. A/B testing de estratégias de retenção

---

8. Resultados Alcançados

Entregas Técnicas

✅ Pipeline completo Bronze-Silver-Gold implementado
✅ 7.043 registros processados sem perdas
✅ 10 features inteligentes criadas
✅ 14 visualizações analíticas geradas
✅ 4 modelos de Machine Learning treinados
✅ Sistema de score de risco individual operacional
✅ Redução de 80,7% no tamanho dos dados (Parquet)
✅ Documentação completa e código versionado no GitHub

Entregas de Negócio

✅ Identificação de R$ 139.130 em receita mensal em risco
✅ 808 clientes priorizados para intervenção
✅ 4 segmentos de clientes mapeados
✅ Janela crítica de 12 meses identificada
✅ ROI potencial: 3,3x na priorização de campanhas

---

9. Próximos Passos e Evolução

Embora o pipeline atual esteja completo e funcional, já identificamos oportunidades de evolução:

Escalabilidade

- **Cloud Computing:** Migração para AWS, Azure ou GCP
- **Processamento em Tempo Real:** Implementação de Apache Kafka para streaming
- **Orquestração Automatizada:** Uso do Apache Airflow para agendamento

Análises Avançadas

- **Deep Learning:** Redes neurais para padrões não-lineares
- **NLP:** Análise de sentimento em tickets de suporte
- **Séries Temporais:** Previsão de tendências futuras de churn

Operacionalização

- **Dashboards Interativos:** Power BI ou Tableau para stakeholders
- **API de Predição:** Endpoint REST para score de risco em tempo real
- **Sistema de Alertas:** Notificações automáticas para clientes de alto risco

---

10. Conclusão

Desenvolvemos com sucesso um **pipeline completo de Big Data** que transforma dados brutos em insights acionáveis e modelos preditivos. Nossa arquitetura Medallion garante qualidade e rastreabilidade em cada etapa, enquanto as técnicas de Machine Learning proporcionam capacidade preditiva com **AUC-ROC de 0,835**.

O projeto demonstra domínio técnico em:

- Engenharia de dados (ETL, Medallion Architecture)
- Análise estatística (testes de hipóteses, correlações)
- Machine Learning (classificação, regressão, clustering)
- Visualização de dados (14 gráficos analíticos)
- Tradução de dados em valor de negócio

Status atual: Pipeline totalmente implementado, testado e operacional. Modelos prontos para deploy e uso em produção.

Arquivos Gerados

Dados

- `/dados/bronze/` - Dados brutos originais
- `/dados/silver/` - Dados transformados (Parquet)
- `/dados/gold/` - Datasets agregados e scores de risco

Visualizações (14 arquivos PNG)

- `viz_01_distribuicao_churn.png`
- `viz_02_churn_por_contrato.png`
- `viz_03_analise_tenure.png`
- `viz_04_charges_vs_churn.png`
- `viz_05_impacto_servicos.png`
- `viz_06_correlacao.png`

- `viz_07_satisfacao_churn.png`
- `viz_08_modelos_ml.png`
- `viz_09_matrizes_confusao.png`
- `viz_10_testes_estatisticos.png`
- `viz_11_regressao_linear.png`
- `viz_12_clustering.png`
- `viz_13_analise_sobrevivencia.png`
- `viz_14_score_risco_lift.png`
- `dashboard_final.png`

Modelos e Outputs

- `scores_risco_clientes.csv` - Scores individuais por cliente
- `relatorio_avancado.txt` - Relatório técnico completo
- Modelos treinados (Pickle/Joblib) - Prontos para deploy

Repositório GitHub: [Link do repositório]
Documentação Completa: README.md no repositório

Documento gerado em 30 de Novembro de 2024
Fundamentos de Big Data - Projeto Final