Telecom - Churn e Agente de Tickets

Índice

Introdução

A empresa de telecom X nos procurou para criar uma pipeline de MLOps aplicando boas práticas de desenvolvimento e dados ao modelo de machine learning de Churn existente. Além disso, precisa da implementação de um modelo de IA Generativa para atender ou resolver chamados (tickets) repetidos, gerando maior eficiência nos atendimentos.

Assistente LLM de Suporte Técnico (Troubleshooting & Retenção)

Telecom é o setor com **maior volume de tickets repetitivos**:

- Problemas de conexão
- Lentidão na internet
- Reset do modem
- Dúvidas sobre fatura
- Mudança de plano
- Portabilidade
- Cancelamento (churn prevention)

Hoje, TODAS as grandes teles estão investindo nisso:

- Vivo → Aura (LLM interno)
- TIM → Assistente baseado em modelos generativos
- Claro → Atendimento baseado em IA
- Verizon → AI-powered troubleshooting
- Vodafone → TOBi usando LLM + RAG

Solução

Um **assistente LLM corporativo para telecom**, capaz de:

1. **Responder dúvidas técnicas do cliente**
2. **Consultar dados do usuário** (RAG + APIs realistas mockadas)
3. **Executar instruções de troubleshooting**
4. **Diagnosticar causa raiz (IA + regras)**
5. **Avaliar risco de churn** (usa seu modelo ML existente)
6. **Recomendar ofertas personalizadas**
7. **Registrar logs para supervisão humana (LLMOps)**

Arquitetura recomendada: MLOps (Modelo de churn)

Stack 100% Google Cloud

- **Cloud Storage** → Bronze, Silver, Gold datasets
- **BigQuery** → Feature store & análises
- **Vertex AI Pipelines** → Orquestração
- **Vertex AI Training** → Treinamento automatizado
- **Vertex AI Model Registry** → Versionamento
- **Vertex AI Batch Prediction** → Predições de churn
- **Cloud Build / GitHub Actions** → CI/CD
- **Cloud Run / Cloud Functions** → Endpoints
- **Pub/Sub** → disparos periódicos
- **Cloud Scheduler** → execução diária/semanal

Fluxo completo

1. **Ingestão de dados** (BigQuery ou CSV)
2. **Pipeline de data prep** (Vertex AI Pipelines + TFX)
3. **Treinamento** (TF/PyTorch/sklearn)
4. **Avaliação e drift detection**
5. **Deploy automático no Vertex Prediction**
6. **Monitoramento contínuo (Vertex AI Model Monitoring)**

Arquitetura recomendada: LLMOps (Assistente Telecom)

Opções de modelo:

- **Gemini 1.5 Pro** (via Vertex AI)
- **Llama 3 70B** (Vertex Model Garden)
- **Mistral** (Model Garden)

Principais componentes LLMOps que usaremos:

✔ **RAG com Vertex AI Search + BigQuery**
✔ **Prompt templates versionados no GitHub**
✔ **Prompt Weights / Eval Sets**

✔ **Vertex AI Evaluation (automated prompt testing)**
✔ **Vertex AI Agents** (parcial ou total)
✔ **Vertex AI Observability**
✔ **CI/CD para LLMs**
✔ **Fine-tuning ou Adapter-training** (opcional)

Pipeline LLMOps

1. **Ingestão de documentos de telecom** (PDFs, base de conhecimento, FAQ)
2. **Criação da base RAG** (Vertex AI Search / Pinecone no GCP)
3. **Serviço de inferência LLM** (Gemini ou Llama)
4. **Avaliação automática** (testes de regressão semântica)
5. **Governança & Safety**
6. **Deploy como API** (Cloud Run)
7. **Logging e observabilidade do LLM**
8. **CICD do pipeline de prompts + testes**

Roadmap recomendado

Grupo 1 — Planejamento inicial

- Criar repositório GitHub
- Criar board Kanban
- Definir arquitetura final
- Criar documentação do projeto
- Provisionar Google Cloud

Grupo 2 — MLOps: ambiente + ingestão

- Criar buckets
- Criar datasets no BigQuery
- Iniciar pipeline de ingestão
- Estruturar Vertex AI Pipeline (orquestração)

Grupo 3 — MLOps: modelo de churn

- Criar notebook de treinamento
- Criar pipeline de treinamento
- Registrar modelo no Model Registry
- Criar endpoint para predição

Grupo 4 — LLMOps: base RAG

- Criar indexação da base de conhecimento
- Configurar Vertex AI Search
- Criar prompts iniciais

Grupo 5 — LLMOps: Assistente Telecom

- Criar API Cloud Run
- Configurar agentes
- Implementar função de troubleshooting
- Conectar com predições de churn

Grupo 6 — CI/CD + observabilidade

- Configurar GitHub Actions
- Automação para modelos ML
- Automação para prompts e LLM
- Dashboards no Cloud Monitoring
- Testes E2E