Utilizando a regressão logística para classificação de churn em um ambiente de startup

Antonio C. da Silva Júnior
Orientador: Walmes M. Zeviani

Especialização em Data Science e Big Data
Universidade Federal do Paraná
Agosto 2020

Contexto

- Relacionamentos de longo prazo são essenciais para o sucesso econômico das empresas
- Modelos de predição de churn são importantes ferramentas para estratégias de retenção de clientes
- Falta de maturidade nos processos e na infraestrutura analítica em ambientes de startup

Estruturação do conjunto de dados

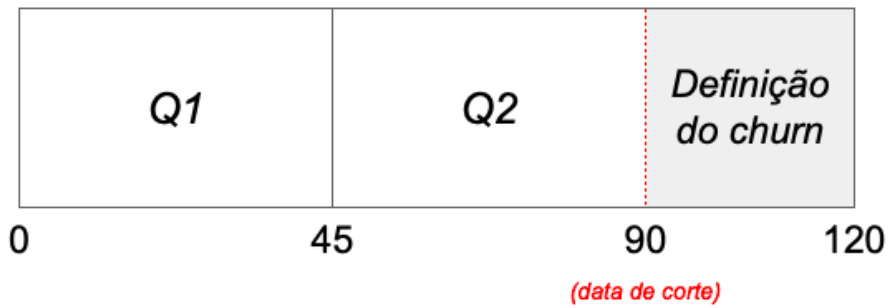
- Base de dados de uma startup brasileira que comercializa uma plataforma para vendas nos grandes marketplaces
- Árduo processo de exploração, preparação e limpeza dos dados
- Neste contexto o vendedor é o cliente da empresa

Definição da variável resposta

- Vendedores com mais de 30 dias inativos foram considerados como churn
- Considerando como inatividade o não acesso à plataforma e a não relização de vendas online

Definição das covariáveis de desempenho

- Data de corte = data da última atividade para os vendedores que deram churn
- Data de corte = data da análise para os vendedores ativos
- Mantidos no conjunto de dados somente os vendedores com pelo menos 90 dias de histórico
- Calculadas diversas métricas em função de $Q1$ e $Q2$
- Estabelecida um medida de desempenho para cada métrica



Inclusão de outras covariáveis

- Qualitativas
- Quantitativas "globais"
- Conjunto de dados final: 35 variáveis e 11131 observações

Comparação dos modelos logísticos

- Modelo 1: Stepwise (19 covariáveis)
- Modelo 2: LASSO (27 covariáveis)
- Modelo 3: LASSO + Stepwise (18 covariáveis)
- Escolhido o modelo 1 (teste da razão da verossimilhança)

Diagnóstico e avaliação

- Análise do comportamento dos resíduos quantílicos aleatorizados
- Análise da curva ROC (AUC: 0,89)
- Escolha do cutoff
- Sensibilidade: 0,84; Especificidade: 0,80
- Acurácia: 0,82 (IC: 0,80 - 0,84)
- Teste de concordância Kappa: 0,64

Obrigado!