Documento de Entendimento

Sprint Session 4 - Ana Health
Grupo - ANA CODA

Leonardo Scarlato, Matheus Aguiar, João Alfredo, Gustavo Antony, Erik Soares, Marcelo Marchetto

O problema que enfrentaremos neste projeto está relacionado ao *churn*, uma métrica de perda de clientes pela empresa. Nesse caso, estamos interessados em observar quais os tipos de clientes que são mais propícios a cancelar ou desistir do plano e calcular qual será a taxa de clientes que cancelaram o seu serviço de assinatura com a Ana Health.

No dia 21 de novembro de 2023, realizamos uma conversa com os membros da *Ana Health*™, nessa conversa, envolvendo todos os grupos da sala de aula ficou evidente que há mais a ser discutido, como por exemplo o fato de que, durante o processo de onboarding, o não prosseguimento com a assinatura também deve ser considerado na nossa análise. Além disso, nossos dados trazem algumas situações onde há um registro de alguém cancelando o plano e voltando a utilizar os serviços da empresa futuramente, assim como funcionários que perderam o benefício da empresa mas continuaram como pessoa física (eles aparecem como tendo cancelado o plano).

Portanto, o principal desafio do nosso projeto é verificar quais clientes cancelaram o serviço, e guardaremos essa informação em uma nova coluna específica para que possamos utilizar essa série como nosso *target*, ou seja, o que queremos prever em nosso modelo de *machine learning* e assim ser possível identificar e prever churn futuro.

Desafios envolvendo o DataSet

O DataSet possui 73 colunas, com muitas informações, indo desde a data de início do contrato até a quantidade de mensagens da equipe que foram direcionadas ao usuário, além de 1202 linhas, cada linha pode ser um usuário (pessoa física) ou uma empresa. Ademais, percebemos, na nossa primeira análise, que muitas colunas contam com valores nulos, e outras com diversos valores para a mesma célula. Isso é provavelmente o resultado de uma junção de diferentes tabelas, problemas esses que teremos que lidar para que não atrapalhe o nosso algoritmo final. Além disso, algumas colunas aparentam estar com o tipo de dado diferente do que é esperado, como por exemplo o status de casado que é um número inteiro.