Documento de Entendimento

Sprint Session 4 - Ana Health
Grupo - ANA CODA

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

A empresa parceira nessa Sprint Session é a Ana Health, uma startup no ramo da saúde que funciona como uma clínica online. Ela atende tanto pessoas físicas quanto empresas, agindo como benefício nesse último caso. A empresa tem aproximadamente dois (2) anos de existência, sendo fundada em 2021.

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.

Desafios envolvendo o DataSet

O DataSet possui 73 colunas e 1202 amostras (linhas), com informações que vão desde a data de início do contrato até a quantidade de mensagens da equipe que foram direcionadas ao usuário. Cada linha de usuário pode ser uma pessoa física, pessoa jurídica ou acolhimento desemprego.

Ademais, percebemos 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), ou valores binários em colunas que representam quantidades (como é o caso da coluna de datas de atendimentos médicos).

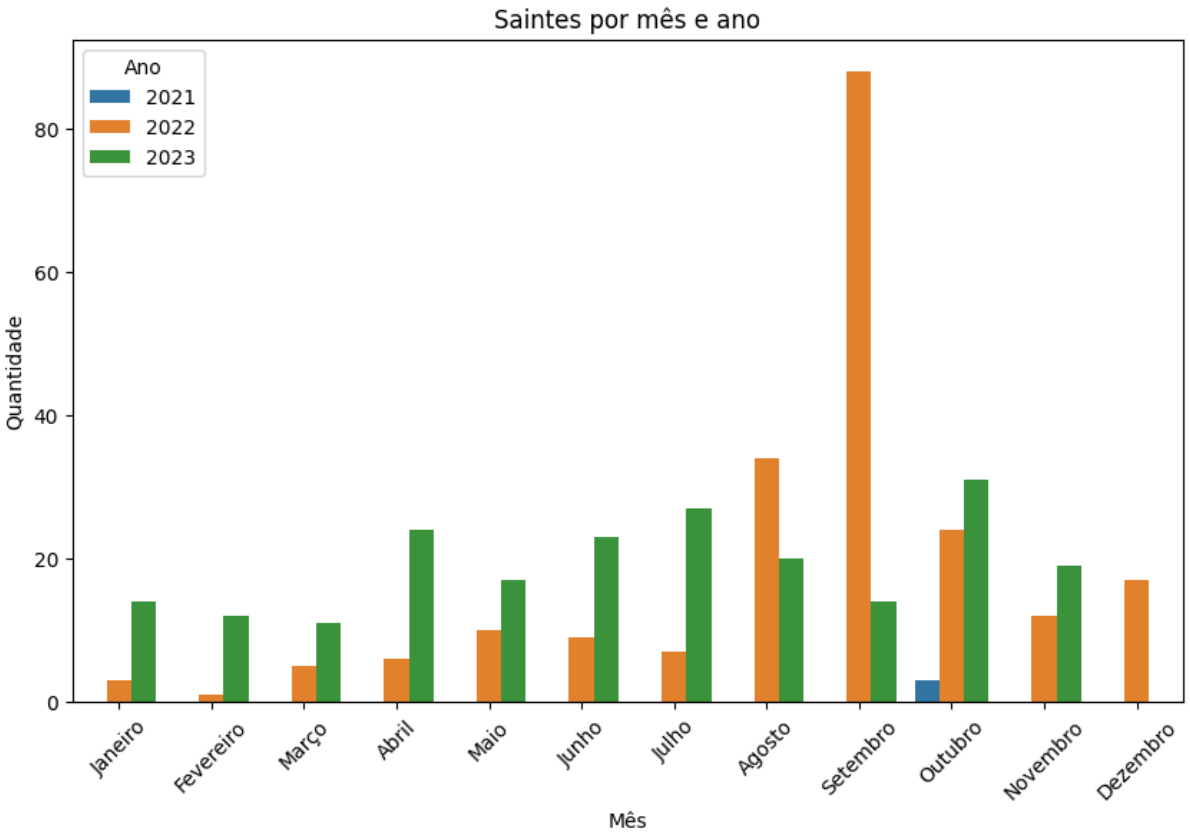
Quando tratamos das pessoas físicas, na maioria das vezes, não observamos o problema relacionado a diversos valores para uma mesma célula. As únicas, possíveis exceções, são relacionadas a notas do teste de WHOQOL (teste de qualidade de vida da World Health Organization), onde uma pessoa pode fazer o mesmo teste mais de uma vez, portanto, tendo mais de uma nota.

Em relação a empresas, podemos diferenciá-las de duas principais maneiras: o valor da célula de *id_org* náo é nulo e há mais de um valor (separado por ponto e vírgula) na coluna de *id_stage*.

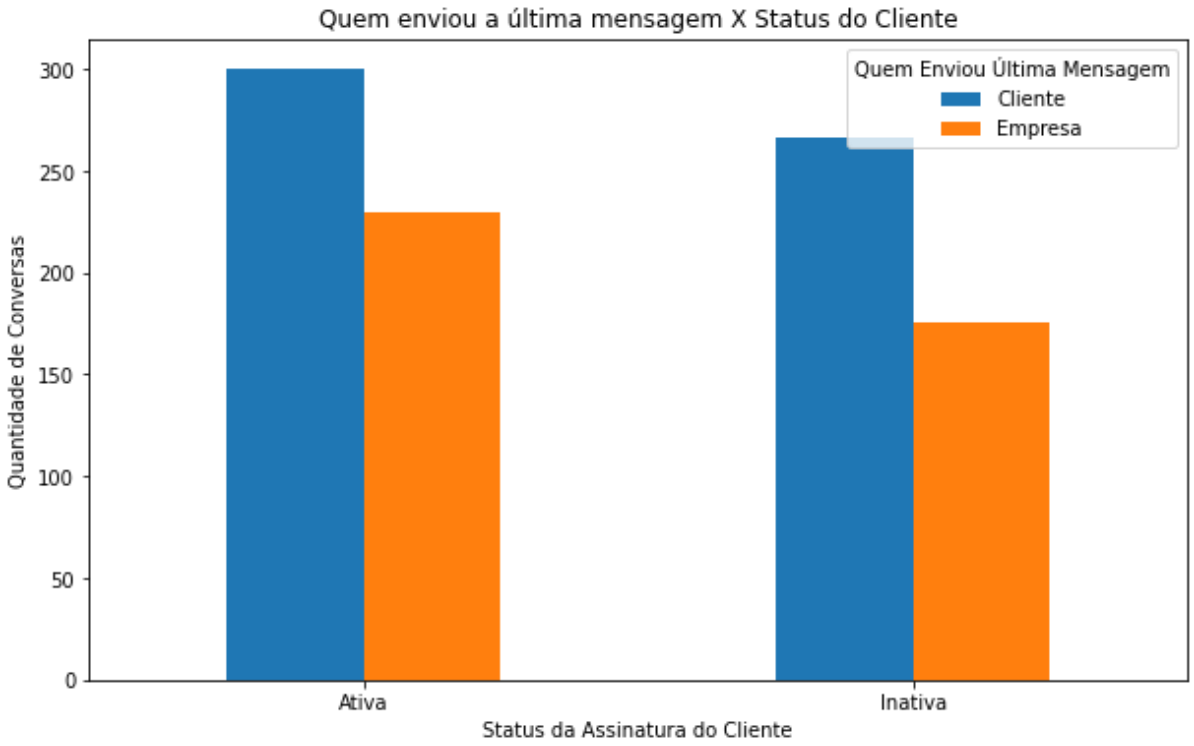
Quatro colunas principais nos indicam possível cancelamento. A primeira é a *contract_end_date*, onde a presença de uma data significa que houve um cancelamento de contrato por parte do cliente. Esse cancelamento pode ser definitivo ou mudança de plano. O status da assinatura nos diz se a assinatura do cliente está em atividade (won) ou se foi interrompida (lost). Há também o *status* do processo de onboarding, que nos diz se o processo foi concluído ou não (algo importante já que também são contemplados como perda de cliente). Por último, a coluna de *lost_time* nos traz informações sobre as datas de cancelamento do cliente, uma para cada assinatura que foi cancelada.

Além dessas, temos também que considerar possíveis faltas de sessões (psicoterapia, atendimento médico e acolhimento) e seu números, se a última mensagem entre o cliente e a equipe de saúde foi feita pela equipe de saúde (na situação onde o cliente for o último a ter enviado a mensagem isso pode indicar uma falha na hora de atender ao pedido), tempo de resposta médio da equipe de saúde e as médias dos teste de WHOQOL (além das diferenças entre testes consecutivos).

Durante a nossa análise exploratória criamos dois gráficos que nos ajudam a entender o comportamento dos clientes da Ana Health. Sendo eles um gráfico que mostra quantos clientes cancelaram o plano a cada mês, e outro que procura uma relação do status da assinatura com o fato do cliente ter sido respondido pela empresa na sua última mensagem ou não.



No gráfico acima, podemos concluir algumas coisas. A mais gritante é a falta de saíntes (pessoas que cancelaram o plano) no ano de 2021. Outra coisa é que no mês de setembro de 2022 houve um aumento expressivo no número de pessoas que cancelaram o seu plano com a Ana Health. Isso já foi explicado durante uma das conversas como a saída de uma empresa parceira. Apesar disso, nos demais meses não observamos uma divergência grande no número de saintes nesses períodos. O ano atual apresenta uma taxa de saíntes mais consistente quando comparada aos demais, com maiores picos em outubro, julho e abril.



Ao propor a análise do gráfico acima acreditávamos que havia a existência de uma correlação entre as mensagens dos usuários serem respondidas e a permanência deles no plano. No entanto, observamos que não é possível chegar nessa conclusão com base nos dados que possuímos por alguns motivos. Ambos os possíveis status do cliente tem uma distribuição parecida, além de que não conseguimos inferir o conteúdo dessa última mensagem. Dessa forma, no momento não encontramos evidências que indiquem que o fato de ter tido sua última mensagem respondida ou não seja um dado relevante para dizermos se o usuário vai cancelar o seu plano.

Descrição dos modelos analisados

Pré-processamento:

Antes da criação dos modelos, foi necessário tratar os dados recebidos pois esses estavam em formatos, e com valores que atrapalhariam o nosso modelo. Começamos selecionando algumas colunas que seria necessário excluir (a maioria por não estarem suficientemente preenchidas ou por apresentarem informações que indicam explicitamente qual é valor da coluna alvo como por exemplo a feature "***lost_reason***" que só tem valor quando alguém já não assina mais a Ana Health), as colunas são: 'state', 'city', 'postal_code', 'id_person_recommendation', 'Recebe Comunicados?', 'Interesses', 'Pontos de Atenção', 'id_stage', 'id_org', 'status.1', 'activities_count', 'Datas Atendimento Médico',

'Datas Acolhimento', 'Datas Psicoterapia','Qde Prescrições', 'Datas Prescrição', 'Qde Respostas WHOQOL', 'id_person', 'contract_start_date', 'contract_end_date', 'id_continuity_pf', 'Canal de Preferência', 'status', 'lost_time', 'add_time', 'id_label' , 'won_time' , 'lost_time.1', 'lost_reason', 'lost_reason.1', 'Qde Atendimento Médico','Faltas Atendimento Médico', 'Qde Atendimentos Acolhimento', 'Faltas Acolhimento', 'Qde Psicoterapia', 'Faltas Psicoterapia', 'Data Última Ligações Outbound', 'Data Última Ligações Inbound', 'Qde Total de Faturas Pagas após Vencimento', 'Qde Perfis de Pagamento Inativos', 'Tempo até Sair', 'Valor Médio da Mensalidade', 'Qde Total de Faturas', 'Problemas Aberto.

Após isso, alteramos as colunas que possuíam mais de um valor para cada linha e selecionamos apenas o último valor (o mais recente ou o único interessante), também calculamos a média dos questionários WHOQOL e preenchemos os valores faltantes pela média das outras médias calculadas. Em sequência convertemos todas as colunas de data para formato datetime

Colunas numéricas:

Em relação às colunas numéricas, muitos dos valores foram transformados. Um exemplo disso são os valores dos testes de WHOQOL (*World Health Organization Quality of Life*) que vieram com múltiplos valores, e foram simplificados para média. Outras, como a ***id_gender***, ***id_marrital_status*** e ***Método de Pagamento*** são categóricas apesar de terem números como valores, portanto, foram abordadas como categóricas.

Outras transformações foram necessárias. Um exemplo foi o que foi feito na coluna de ***notes_count***, onde optamos por retirar valores menores que 7 por conta da sua baixa frequência. Outro é o que foi feito com valores nulos nas colunas ***Mensagens Inbound***, ***Mensagens Outbound***, ***Ligações Inbound***, ***Ligações Outbound***, ***Qde Total de Tentativas de Cobrança***, ***Valor Total Inadimplência***, ***Tempo Última Mensagem Inbound*** e ***Tempo Última Mensagem Outbound*** onde preenchemos valores nulos com 0. Uma coluna mais específica foi a de ***Qde Total de Faturas Inadimplentes***, que foi preenchida da mesma maneira mas foi transformada em booleana, contendo valores verdadeiro se há uma fatura inadimplente e falso se não há.

Colunas categóricas:

O dataset possui um grande número de colunas categóricas. Nelas, geralmente temos números representando classes. Isso não seria um problema se essas classes fossem numericamente relevantes, mas, já que um número maior não é um indicativo de qualidade ou de preço (por exemplo), optamos por uma abordagem de OneHot Encoding mesmo nessas situações onde os dados já são representados por números. As colunas que foram transformadas foram: ***id_gender***, ***id_marital_status***, ***id_health_plan***, **notes_count***,* ***Métodos de Pagamento***, ***Qde Total de Faturas Inadimpletes*** e uma coluna criada na análise que é ***Quem Enviou Última Mensagem***.

Na maioria desses casos, optamos por não colocar o nome de cada ID, apenas colocando o identificador como string e passando os dados pela função get_dummies(), que é responsável pelo OneHotEncoding dos dados. Outras análises foram necessárias, como no ***id_gender*** onde optamos por deixar apenas homens e mulheres no dataset por conta da baixa frequência de outras classes.

Além disso, transformamos a coluna de ***lost_reason*** para tratar as amostras de usuários que ainda estavam ativos.

As colunas de ***status*** (uma referente ao processo de onboarding e outra a assinatura) foram tratadas para incluir apenas o último status do cliente (que é o mais recente cronologicamente).

Além dessas colunas, optamos por criar uma nova coluna categórica que referencia se o cliente tem problema em aberto ou não. Essa coluna foi criada a partir da coluna de problema em aberto, mas tentamos simplificar essa ao máximo depois de ver que a grande maioria das amostras na tabela eram únicas (impossibilitando um OneHot).

Das colunas que foram droppadas, as categóricas são: state, city, postal_code, id_person_reccomendation, Recebe Comunicações?, Interesses, Pontos de Atenção, id_stage, id_org e status (processo de onboarding). Algumas das colunas estavam completamente nulas, outras estavam muito relacionadas ao target do modelo principal (de se o usuário irá sair no mês seguinte) e outras tinham uma correlação muito grande com outras features.

Dataframes e modelos:

Temos dois dataframes que pretendemos utilizar. O primeiro, é referente a análise de quem sai no próximo mês. O segundo, é uma tentativa de prever quanto tempo até o usuário sair.

DataSet 1 - Saída de Clientes no Próximo Mës :

Nele, temos uma coluna de ‘target’ que é binária (valores verdadeiros indicam que a pessoa saiu no próxima mës). As colunas que contemplamos tem relação com o perfil do cliente, como idade, sexo, estado civil, etc. A maioria dos dados já passaram pelo script_dataframe (e a função `tratamento`) para adição de colunas novas, tratamento de datas e criação de dummies.

DataSet 2 - Tempo até o Cliente Sair

O processo de criação foi parecido com o anterior, mas aqui temos uma nova coluna de ‘target’ que é contínua (valores indicam quantos dias até o cliente sair). As colunas que contemplamos tem relação novamente com o perfil do cliente e suas características pessoais. A maioria dos dados também passaram pelo script_dataframe (e a função `tratamento`) para adição de colunas novas, tratamento de datas e criação de dummies. A diferença é que contemplamos as colunas de faltas de atendimentos e consultas.

Portanto, como citado acima, decidimos seguir duas linhas de predição. Para fazer a classificação se um indivíduo vai ou não cancelar o plano da Ana Health, utilizamos os modelos de classificação Random Forest Classifier e Decision Tree Classifier, ambos obtiveram o valor de 100% de acurácia, visto que o nosso Dataframe possui dados sem mudança ao longo do tempo, ou seja uma vez que a pessoa tenha o estado won, ela sempre terá o estado won, esta análise será modificada visto que só percebemos o erro do Dataframe no dia 30/11. Já para o nosso modelo de regressão, utilizamos o Random Forest Regressor obtendo um erro de aproximadamente 50 dias na estimativa de tempo que o cliente permanece no plano. Para fins de análise de coerência do nosso modelo, fizemos uma “predição” com os dados de clientes que ainda estão ativos considerando o tempo total de permanência no plano até o momento atual. Dessa maneira, obtivemos um erro de aproximadamente 100 dias com o modelo prevendo valores um pouco menores do que os verdadeiros, o que possibilita a ação prévia que busca recuperar clientes à beira de sair.