Entendimento do problema e dataset - Entrega 1

A empresa Ana Health, que nos passou o problema, é uma empresa jovem que se difere das demais do ramo de saúde por sua proposta de flexibilizar o atendimento entre doutor e paciente, oferecendo consultas e cuidado proativo através de variadas plataformas digitais e horários de atendimento. Além disso, é uma empresa que utiliza análise de dados dos usuários para que eles recebam conselhos para se manterem saudáveis, oferecendo um dinamismo e um grau de personalização que planos de saúde convencionais não têm.

O problema proposto para nós resolvermos é o diagnóstico dos motivos pelos quais os clientes podem cancelar seu serviço, ou seja, minimizar o Churn, para que a empresa conseguisse melhorar a retenção de clientes. Sabendo e tendo uma visão clara dos motivos que podem levar clientes a cancelarem um plano, a empresa consegue trabalhar nesses motivos e ter um crescimento sustentável, uma vez que manter clientes é mais barato e mais rentável do que apenas conseguir novos.

Para desenvolver uma solução para esse problema, utilizaremos técnicas de machine learning e inteligência artificial, aprendidas ao longo do semestre. Através delas, poderemos filtrar e tratar dados, identificar padrões, fazer previsões, e disponibilizar as informações mais importantes através de uma aplicação, com fácil visualização.

O dataset Ana Health_Tabela Modelo Previsão Churn é dividido em 73 colunas que possuem 1204 clientes, sendo esses 1204 dividido em 289 pessoas físicas, 887 organizações (836 pessoas jurídicas) e 19 acolhimento desemprego. Além disso, os clientes são majoritariamente do estado de São Paulo.

Analisando o dataset vimos que, a data de início do primeiro contrato e de término do último contrato de assinatura do benefício da Ana Health foram:

- 2020: 4 aberturas e não possui informação de encerramentos.
- 2021: 176 aberturas e 34 encerramentos.
- 2022: 508 aberturas e 283 encerramentos.
- 2023: 514 aberturas e 247 encerramentos.

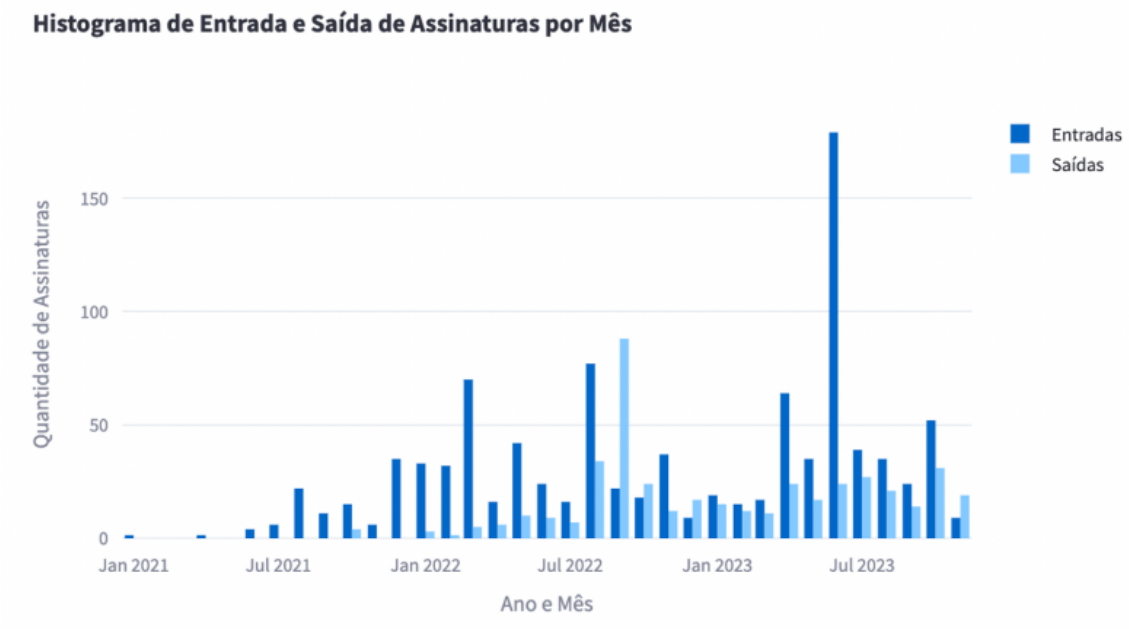
Sendo contrato mais antigo aberto: 16/12/2020 e o contrato mais recente aberto: 08/11/2023. Já o contrato mais antigo encerrado foi : 19/02/2021 e o contrato mais recente encerrado foi: 08/11/2023. Além disso, 347 pessoas deram continuidade ao serviço como pessoa física após o cancelamento por parte da empresa.

Ademais, em relação aos gêneros dos clientes, 602 clientes são do gênero masculino (64), 571 são do gênero feminino (63), 6 são outros gêneros e 24 estão sem informação (NULL). Também descobrimos que a idade média dos clientes é de 32 anos.

Sobre o funil de vendas, o status por assinatura: Cancelada (lost), Ativa (won), analisamos que a Ana Health teve 592 assinaturas ativas e 520 assinaturas canceladas. Sendo os motivos mais recorrentes de cancelamento: Cliente não quis seguir com a Ana, a empresa cancelou o benefício da Ana ou o cliente precisou cortar custos.

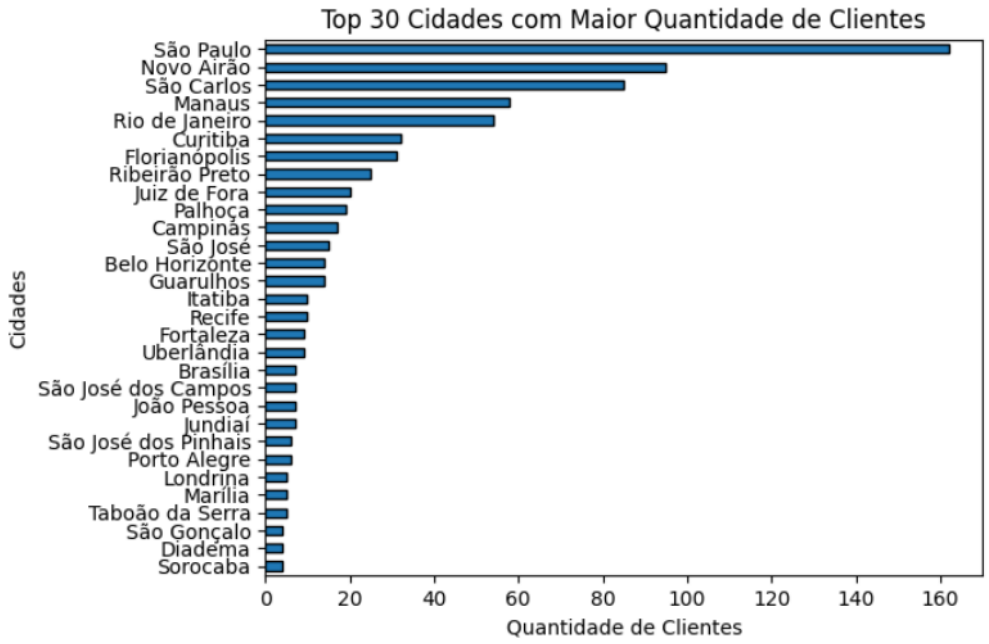
Gráficos da análise exploratória: - Entrega 2

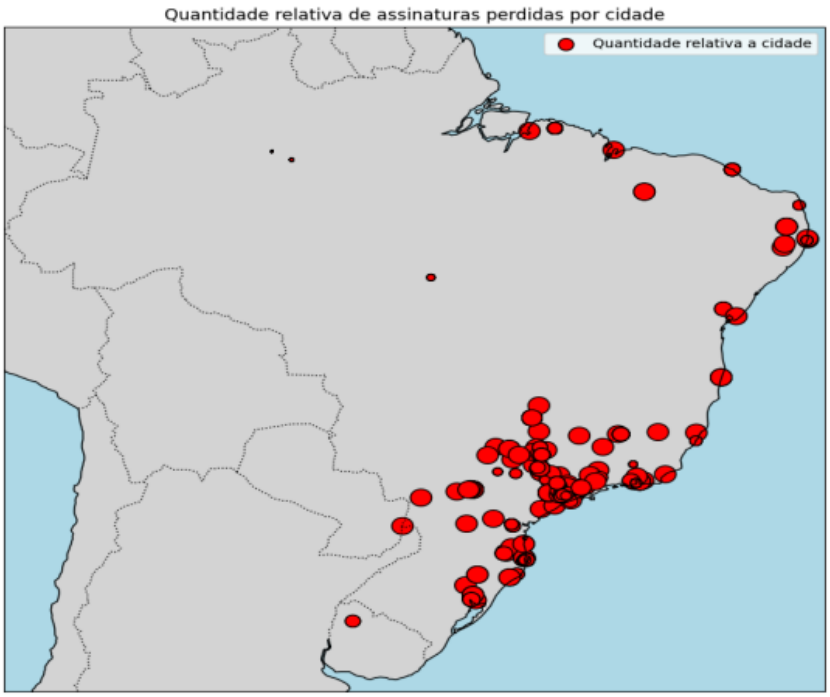
Gráfico 1:



A imagem representa a quantidade de entradas e saídas ao longo dos meses. Sendo que Entradas significa: início do primeiro contrato de assinatura e Saídas: término do último contrato de assinatura. Dessa forma pelo histograma é possível analisar padrões interessantes, como uma grande entrada de novos contratos de assinatura em 2022-02, 2022-07, 2022-03, 2023-05, 2023- 09. Assim como padrões de elevados de saída em datas específicas como em 2022-08. Uma possibilidade interessante seria comparar essas datas atípicas com eventos externos, como entrada ou saída de empresas com clientes.

Gráfico 2:





Este gráfico ilustra as perdas de assinatura em diferentes cidades do Brasil ao longo do período de 2021 a 2023. Os círculos vermelhos no gráfico representam indivíduos que cancelaram suas assinaturas. A dimensão dos círculos é proporcional ao tamanho populacional de cada cidade. Em outras palavras, se, por exemplo, houve 10 cancelamentos na cidade de São Paulo, o tamanho do círculo correspondente será calculado como 10 dividido pela população de São Paulo. Vale ressaltar que a maioria dos cancelamentos ocorreu nas regiões Sul e Sudeste do país.

Modelo Preditivo supervisionado – Entrega 3

Contexto

Para desenvolver um modelo preditivo que fosse útil para AnaHealth, era necessário que com as suas predições fosse possível gerar insights e análises sobre a saída de clientes da plataforma. Para que assim o objetivo de Reduzir o churn em um ponto percentual em até 3 meses fosse alcançado.

Ideia do Modelo

Decidimos utilizar o tempo de permanência na plataforma como target. Para que a empresa obtivesse a previsão de quantos dias cada usuário seria assinante do plano. Com esta previsão a Ana Health poderia agir sobre usuários que estariam próximos de cancelar, com o intuito de manter esses usuários na plataforma e assim reduzindo o churn.

Pré Processamento e Feature Engineering

Antes do desenvolvimento do modelo, manipulamos os dados.

Na etapa de pré processamento trabalhamos com o objetivo específico de lidar com valores nulos. Neste sentido, adotamos estratégias como descartar colunas com muitos valores nulos, consideradas irrelevantes para o modelo, ou preencher esses valores nulos com o número 0.

Além disso, tivemos a etapa de feature engineering, onde focamos em tratar colunas com valores categóricos, aplicando técnicas de encoding e outros métodos para transformar esses dados em formatos válidos, e para que eles fossem mais relevantes em informações. Esse trabalho foi crucial para assegurar que os modelos pudessem interpretar e utilizar boa parte das informações disponíveis no dataset de forma otimizada.

Toda a documentação detalhada com nossas abordagens em cada uma das colunas, onde encontrar os trechos de códigos e como está sendo replicado o processamento está no Documento: Documentação-Manipulação-de-dados.pdf, encontrado na pasta doc a partir da raiz de nosso repositório.

Dados para treinamento

O nosso dataset manipulado foi filtrado para incluir somente as linhas que representam assinaturas canceladas, uma vez que não queremos treinar nosso modelo de tempo de permanência. Dividimos o conjunto de dados em variáveis independentes e dependentes, sendo a 'stay_time' a variável dependente que indica o tempo em dias que um usuário permaneceu ativo na plataforma antes de cancelar a inscrição. Para assegurar a reprodutibilidade e a validade dos resultados, os dados foram divididos em conjuntos de treino e teste, com 25% destinados ao teste, utilizando uma seed aleatória.

Implementamos a técnica de Grid Search com diversos algoritmos e parâmetros para identificar o de melhor desempenho. Entre os métodos testados estava a Regressão Linear, usando features polinomiais e padronização dos dados. No nosso Grid Search, variamos os graus dessas features polinomiais, testando 1, 2 e 3. O desempenho do modelo foi avaliado principalmente através do Mean Squared Error (MSE) e do R2 score.

Além da Regressão Linear, utilizamos o Random Forest Regressor e o Support Vector Regressor, cada um submetido a um Grid Search para determinar os melhores parâmetros e alcançar o melhor desempenho. Os resultados obtidos foram:

Random Forest Regressor:

- RMSE: 96.99
- R2: 0.25312721913815717
- EVS: 0.32890013577762933
- Erro percentual médio: 15.93%

Support Vector Regressor:

- RMSE: 85.63
- R2: 0.4178046750004776
- EVS: 0.43484915131508506

- Erro percentual médio: 14.06%

Linear Regression:

- RMSE: 167.07
- R2: -1.2161651340448807
- EVS: -1.1666649285715551
- Erro percentual médio: 27.43%

Utilizamos métricas como RMSE, R2, EVS e o erro percentual médio para avaliação. RMSE significa “erro quadrático médio”, R2 é o coeficiente de determinação e EVS significa “explained variance score”. Essa métrica indica quão bem o modelo consegue capturar variação nos dados de teste. A fórmula dele é igual a:

$$EVS(y, \hat{y}) = 1 - \frac{Var(y-\hat{y})}{Var(y)}$$

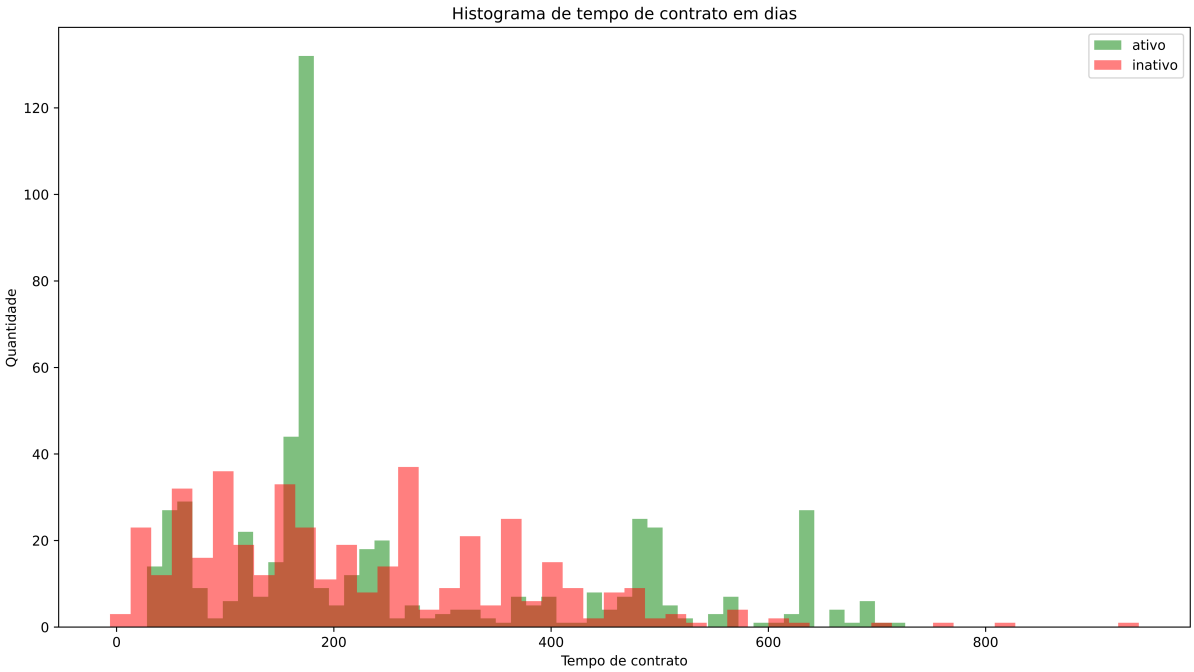
Fonte: Explained variance score as a risk metric - Mastering Python for Finance - Second Edition [Book] (oreilly.com)

Comparando os modelos, concluímos que o Support Vector Regressor apresentou o melhor desempenho, com os hiperparâmetros 'poly_features__degree': 1, 'svr__C': 1, 'svr__epsilon': 0.1, 'svr__kernel': 'linear'. As métricas incluíram R2 e EVS mais próximos de 1, além dos menores valores de RMSE e erro percentual médio.

Limitações do modelo

Pensávamos que o modelo com a target 'stay_time' agregaria valor a Ana Health, ao prever o tempo de permanência de todos os usuários. Contudo, após a construção do modelo e análise mais profunda dos dados chegamos a conclusão que nosso modelo não deve ser usado para a análise pré cancelamento dos usuários.

Considerando que possam existir dois grupos diferentes de clientes, um grupo que tende a permanecer com a assinatura da Ana Health por um longo período, outro grupo que tende a cancelar a assinatura da Ana Health em um curto período. Nosso modelo fica incoerente para a previsão de permanência de usuários ainda ativos, uma vez que por a Ana Health ser uma empresa nova, ainda não deu tempo para que usuários que tendem a ficar um longo período saíssem da plataforma. Por esse motivo nosso modelo só teria contato com clientes que tem a tendência de sair cedo da plataforma



A maior parte das pessoas que saíram, como mostrado no gráfico, ficou por pouco tempo na plataforma. Como o modelo é treinado nesses dados, isso vai se refletir na hora de prever quanto tempo que pessoas que estão na plataforma vão continuar, e a previsão na maioria das vezes vai ser que elas vão sair em pouco tempo, o que pode não refletir a realidade.

Aplicações do modelo

Como nosso modelo deve ser aplicado para reduzir o Churn? Como vimos nas limitações ele não é válido para previsão de permanência de usuários ativos. Nosso modelo deve ser aplicado para prever o tempo de permanência em usuários que já cancelaram o plano, dessa forma será possível identificar casos em que um usuário ficou mais tempo com o plano que o esperado, e casos em que o usuário ficou menos tempo que o esperado. Com base nesse “outliers” é possível identificar possíveis razão e padrões para que um usuário fique mais tempo com o plano, e é possível identificar motivos e fatores pelo qual dado usuário ficou menos tempo na plataforma que o esperado.

Assim recolhendo informação valiosas, sobre a permanência de usuários na plataforma, que podem ser utilizadas para reduzir o churn.