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.

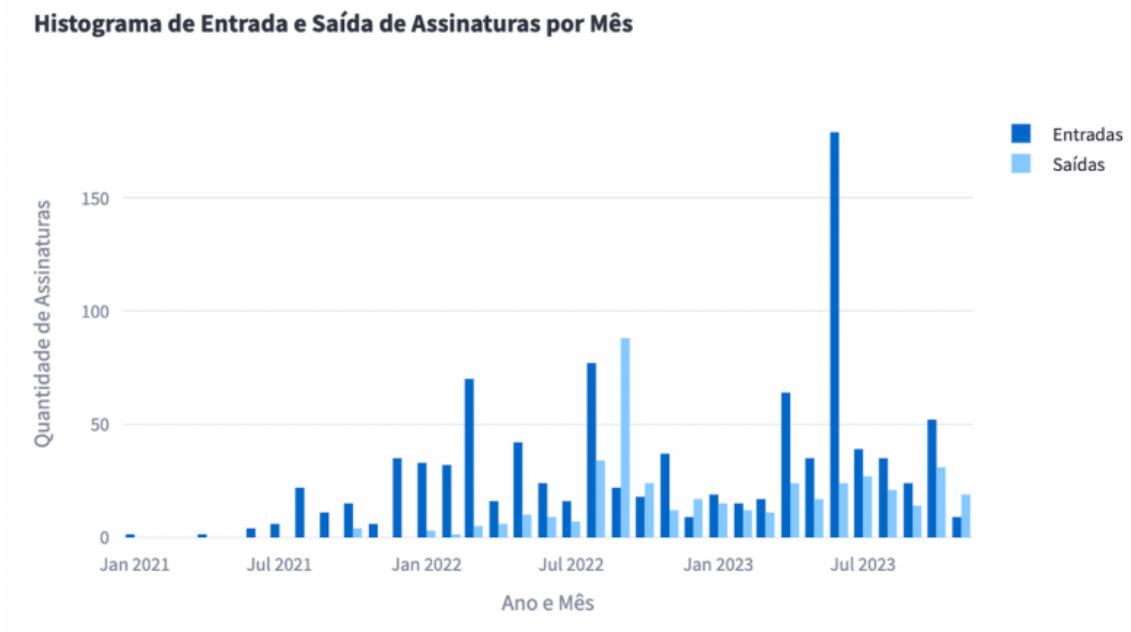
Sendo contrato mais antigo aberto: 16/12/2020 e o contrato mais recente aberto: 08/11/2023. Já o contrato mais antigo encerrado foi : 19/02/2021 e o contrato mais recente encerrado foi: 08/11/2023. Além disso, 347 pessoas deram continuidade ao serviço como pessoa física após o cancelamento por parte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

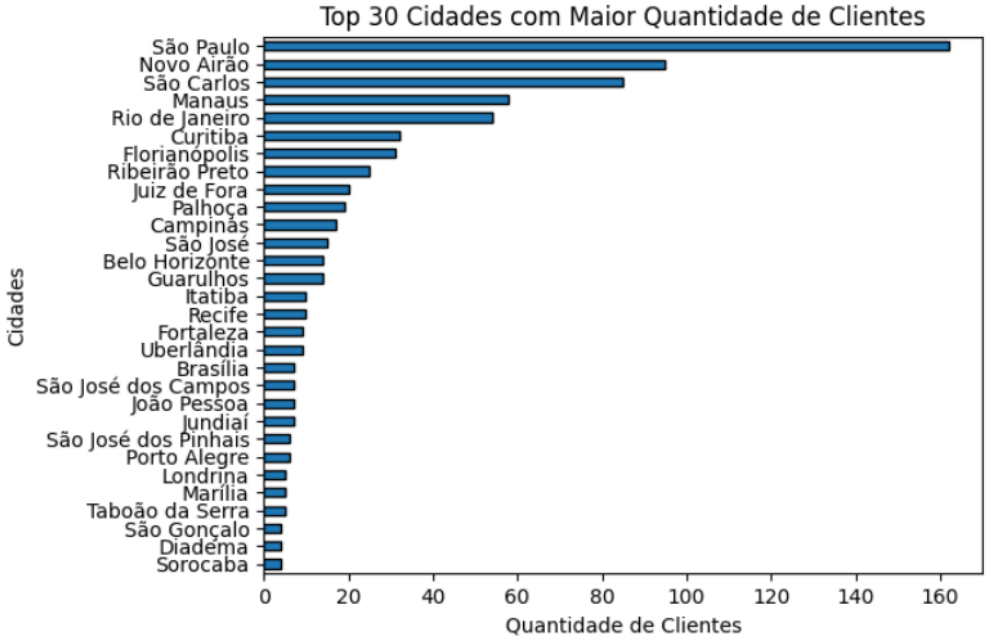
Foram feitos gráficos para ilustrar a análise exploratória, como solicitado na entrega 2, a seguir:

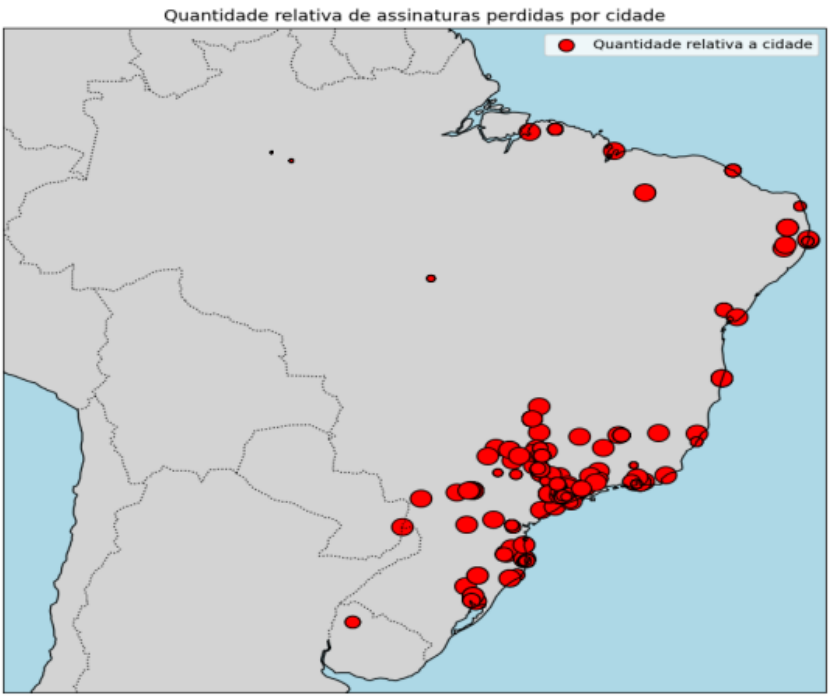
Gráfico 1:



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.

Para a entrega 3, no notebook `modelo_tempo_permanencia.ipynb`, desenvolvemos um modelo preditivo para estimar o tempo de permanência dos usuários na Ana Health. Este processo contou com a utilização de um dataset pré-processado e com feature engineering. Vale destacar que o pré-processamento foi realizado nos notebooks contidos na pasta 'pré-processamento' dentro da pasta 'notebooks', 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, a feature engineering foi realizada nos notebooks localizados na pasta 'notebooks/feature engineering'. Nessa fase, focamos em tratar colunas com valores categóricos, aplicando técnicas de encoding e outros métodos para transformar esses dados em um formato que pudesse ser efetivamente utilizado pelos algoritmos de aprendizado de máquina. Esse trabalho foi crucial para assegurar que os modelos pudessem interpretar e utilizar todas as informações disponíveis no dataset de forma otimizada.

O nosso dataset foi filtrado para incluir somente as linhas que representam assinaturas canceladas, em linha com o foco do nosso modelo. 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: 78.69
- R2: 0.4634757948768473
- EVS: 0.518854857312801
- Erro percentual médio: 27.41%

- **Support Vector Regressor**:
- RMSE: 85.16
- R2: 0.4241861950076372
- EVS: 0.4404396850224982
- Erro percentual médio: 13.98%

- **Linear Regression**:
- RMSE: 166.92
- R2: -1.2121648332607773
- EVS: -1.1650483870435169
- Erro percentual médio: 13.51%

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:

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

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.

Este modelo com a target 'stay_time' é de grande valor para a Ana Health, especialmente no contexto de redução de churn. Ao prever o tempo de permanência dos

usuários, a empresa pode identificar padrões e tendências que indicam um risco maior de cancelamento de assinatura. Com essas informações, a Ana Health pode implementar estratégias proativas para aumentar a retenção de usuários, oferecendo, por exemplo, incentivos ou melhorias personalizadas para aqueles com maior probabilidade de churn. Dessa forma, este modelo não apenas fornece insights valiosos para a tomada de decisão, mas também é uma ferramenta poderosa para melhorar a satisfação do cliente e fortalecer a fidelidade à marca.