MODELO PREDITIVO DE CHURN DE ENTREGADORES

EMPRESA RAPPI

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	-Alysson Cordeiro; -Bruno Moitinho Leão; -Frederico Schur; -Israel Carvalho; -Luiz Carlos da Silva Júnior; -Stefano Tosi Butori.	1.0	Criação do documento
09/08/2022	Todos	1.1	Atualização dos tópicos 4.1.1 a 4.1.5
10/08/2022	Todos	1.2	Atualização do tópico 4.2

Sumário

1. Introdução **5**

2. Objetivos e Justificativa **6**

2.1. Objetivos 6

2.2. Justificativa 6

3. Metodologia **7**

3.1. CRISP-DM 7

3.2. Ferramentas 7

3.3. Principais técnicas empregadas 7

4. Desenvolvimento e Resultados **8**

4.1. Compreensão do Problema 8

4.1.1. Contexto da indústria 8

4.1.2. Análise SWOT 11

4.1.3. Planejamento Geral da Solução 12

4.1.4. Value Proposition Canvas 13

4.1.5. Matriz de Riscos 15

4.1.6. Personas 17

4.1.7. Jornadas do Usuário 17

4.2. Compreensão dos Dados 18

4.2.1. Descrição Geral dos Dados 18

4.2.2. Estatística Descritiva dos Dados 20

4.2.3. Descrição da Predição Desejada ("Target") 20

4.3. Preparação dos Dados 21

4.4. Modelagem 22

4.5. Avaliação 23

4.6 Comparação de Modelos 24

5. Conclusões e Recomendações **25**

6. Referências 26

Anexos 27

1. Introdução

Apresente de forma sucinta o parceiro de negócio, seu porte, local, área de atuação e posicionamento no mercado. Maiores detalhes deverão ser descritos na seção 4

Descreva resumidamente o problema a ser resolvido (sem ainda mencionar a solução).

Caso utilize citações ao longo desse documento, consulte a norma ABNT NBR 10520. Sugerimos o uso do sistema autor-data para citações.

2. Objetivos e Justificativa

2.1. Objetivos

Descreva resumidamente os objetivos gerais e específicos do seu parceiro de negócios

2.2. Justificativa

Faça uma breve defesa de sua proposta de solução, escreva sobre seus potenciais, seus benefícios e como ela se diferencia.

3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

3.1. CRISP-DM

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Collaboratory)

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

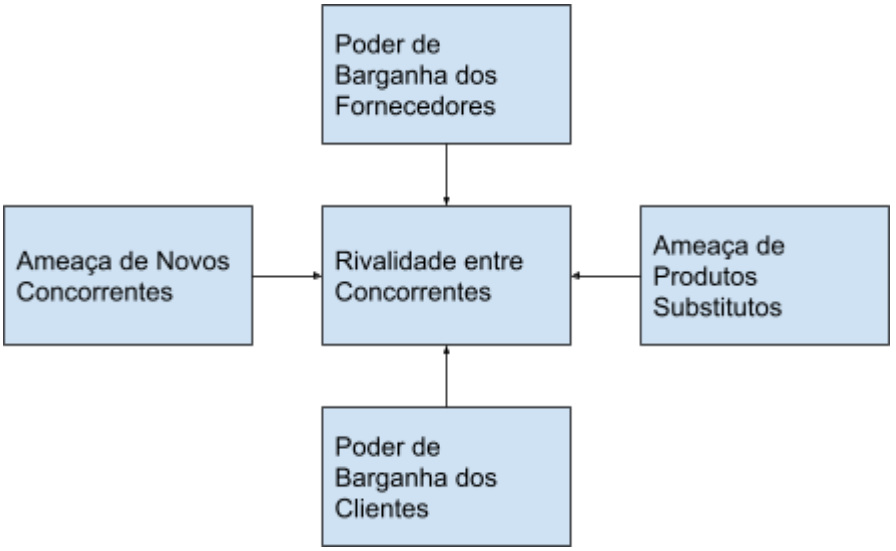
4.1.1. Contexto da indústria

O modelo de negócio da Rappi é do tipo plataforma multilateral, pois conecta estabelecimentos comerciais a entregadores e a clientes que precisam de serviços de entrega (delivery). A plataforma disponibiliza cardápios e menus das mais variadas instalações - restaurantes, lojas de todos os ramos, supermercados, entre outros.

Além do mais, a concorrência da companhia é bastante abrangente, porque, de certa forma, inclui desde toda e qualquer empresa de delivery (especialmente os aplicativos de entregas, como, por exemplo, IFood, Loggi, Glovo e Lalamove) e até os e-commerces gigantes, como a Amazon; esse de uma forma indireta.

Vale ressaltar, ademais, que o delivery é uma das principais tendências do mercado após o isolamento social da pandemia do COVID-19 em 2020. E ele tende apenas se fortalecer. No entanto, as entregas estão deixando de ser focadas apenas em comida para alcançar novos patamares, como medicamentos e outros produtos diversos, por exemplo. Além disso, esse tipo de mercado está optando por uma prática mais sustentável, como, por exemplo, ajudar o consumidor a identificar restaurantes que usam menos plásticos em seus produtos e tornar grande parte dos pedidos neutros em carbono (CO_2) no país.

Utilizaremos a análise das 5 Forças de Porter para entender melhor a indústria de delivery por aplicativo.



Ameaça de Novos Concorrentes:

Durante a pandemia, 89% dos estabelecimentos passaram a utilizar o delivery como uma solução para disponibilizar seus produtos (grande crescimento do mercado). Entretanto, o alto custo operacional gerado pela operação no mercado de delivery acaba dificultando a instalação de novas empresas, além da manutenção de empresas já estabelecidas, como por exemplo o recente caso do Uber Eats (alto custo operacional).

Poder de Barganha dos Clientes:

Por mais que houve um aumento expressivo no setor de delivery impulsionado principalmente pela necessidade das empresas de se adequarem às restrições impostas ao SARS-CoV-2 (coronavírus)e, também, considerando o cenário macroeconômico dos últimos meses, é possível notar um aumento da inflação e, consequentemente, uma diminuição no grau de consumo em diversos áreas da economia, incluindo a de delivery. Logo, é razoável inferir que o poder de barganha dos consumidores nesse cenário é relativamente baixo, visto que diante de um cenário em que há uma alta de preços em toda a economia, os consumidores não conseguirão "barganhar" por preços menores.

Ameaça de Produtos Substitutos:

Não há muitos produtos substitutos para o delivery por aplicativo. A opção alternativa mais evidente seria a entrega efetivada diretamente em contato com o restaurante, eliminando a intermediação via aplicativo. Essa opção não oferece a comodidade de ver o cardápio de opções pelo aplicativo e apresenta o incômodo de ter que pesquisar um restaurante e conversar com um atendente por telefone.

Poder de Barganha dos Fornecedores:

Tendo em vista todo o cenário descrito nos tópicos anteriores, é possível concluir que o poder de barganha dos fornecedores é expressivo, uma vez que eles controlam em qual marketplace estarão disponíveis. Por outro lado, contratos de exclusividade podem restringir a capacidade de escolha dos fornecedores e reduzir o poder de barganha destes.

Rivalidade entre Concorrentes:

A rivalidade entre os concorrentes nesse setor é extremamente alta em consequência do constante esforço para garantirem uma operação de qualidade, além de um grande market share. É comum que empresas desse setor gastem seus recursos de caixa para crescer, e, consequentemente, acabam não dando lucro.

4.1.2. Análise SWOT

Strengths (Forças)	Weakness (Fraquezas)
- Entrega turbo em 10 minutos - Alguns parceiros exclusivos - Parceria com players de outros setores (ex: HBOMax) - Presença substancial na América Latina - Atuação em outros setores, como o financeiro (RappiBank)	- Dependência de entregadores que não são CLT - A função de rastreio da entrega não é consistente - "Churn" elevado de entregadores
Opportunities (Oportunidades)	**Threats (Ameaças)**
- Saída recente de um player relevante, abrindo assim uma nova fatia do mercado - Aumento da demanda por delivery de diversos produtos	- Mercado bastante competitivo - Instabilidade de fatores externos que afetam os entregadores, como preço da gasolina, variações climáticas ou feriados

4.1.3. Planejamento Geral da Solução

Dados disponíveis: A Rappi disponibilizou até o momento uma série de arquivos CSV, com o conteúdo detalhado de forma mais extensiva no tópico 4.2.1 deste documento. Os dados desses arquivos, conforme informações repassadas pelo cliente em entrevista, foram obtidos por meio do aplicativo de delivery.

Esses arquivos incluem informações diversas sobre os entregadores, como a taxa de aceitação dos pedidos, receita mensal, reclamações sobre pedidos, data de criação das contas que sofreram "churn", registro de suspensões e avisos, número de ordens, tempo que o entregador ficou "online", ordens devolvidas e dados gerais.

Solução proposta: Pretendemos desenvolver um modelo preditivo para classificar se um entregador irá permanecer ou sair ("churn") da plataforma Rappi. Como efeito secundário, podemos verificar a probabilidade que o modelo entende que aquele resultado se concretizará e os fatores ("features") que o modelo considera para a classificação geral.

Tipo de tarefa (regressão ou classificação): a predição de "churn" é essencialmente uma tarefa de *classificação binária* - o objetivo é informar se os entregadores irão: (i) permanecer; ou (ii) sair ("churn") da plataforma dentro de um período determinado.

Forma de utilização da solução proposta: a Rappi poderá usar o modelo para classificar se um entregador está propenso a sair da plataforma e, se desejar, dedicar mais esforços para manter esse entregador na plataforma.

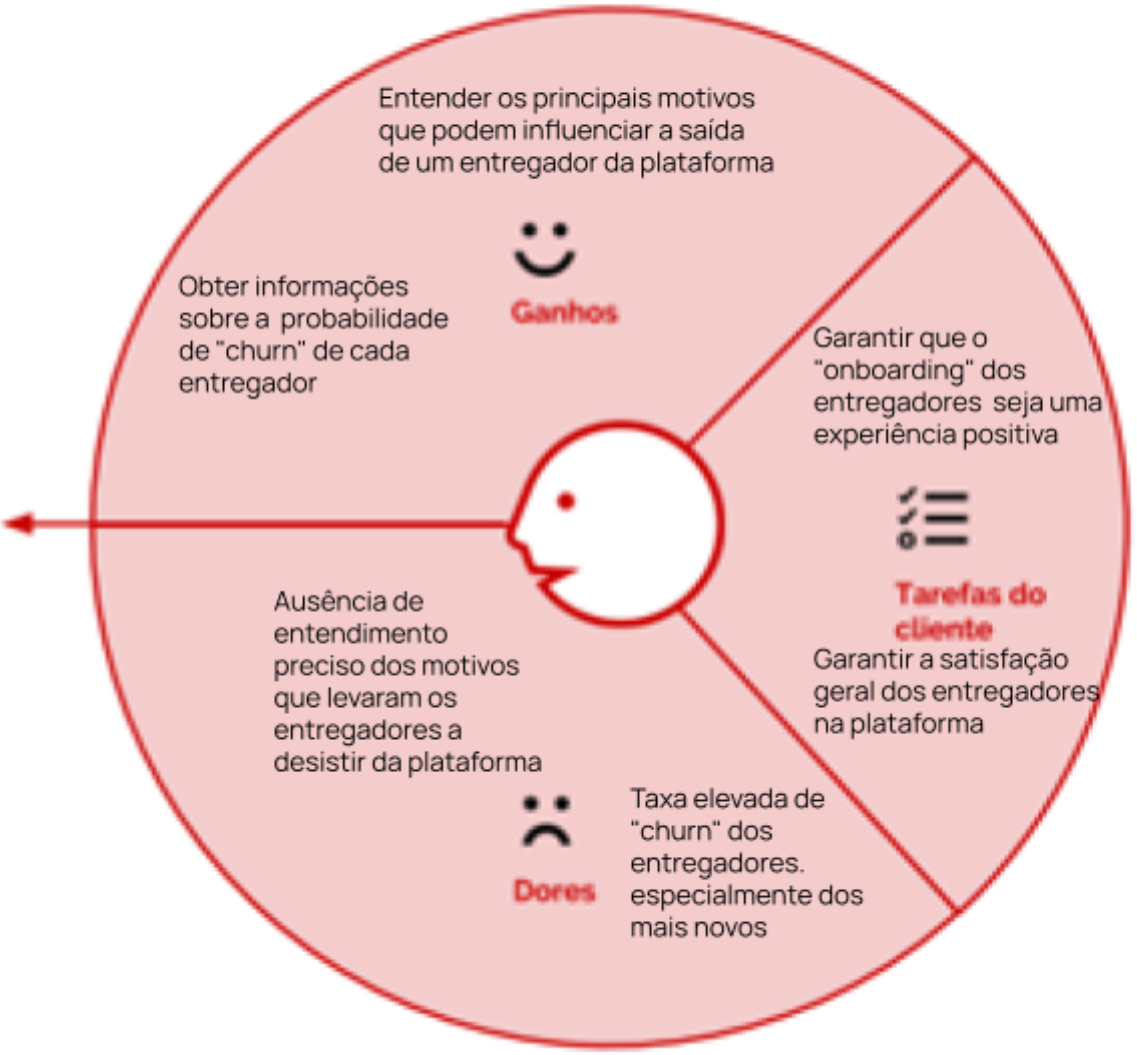
Benefícios trazidos pela solução proposta: o modelo pode permitir uma visualização rápida dos entregadores mais inclinados a parar de usar a plataforma em breve, levar a um maior entendimento dos fatores que podem causar esse evento, facilitar o diálogo e reduzir a taxa de "churn" para um patamar adequado.

Critério de sucesso e métricas de avaliação: entendemos que o modelo deve apresentar certa probabilidade de acerto em suas predições. Para tanto, utilizamos como métricas de avaliação:

A. acurácia ("accuracy"), a razão entre o número de predição corretas e o número total de predições, fornece um indicador de confiabilidade geral do modelo: $\frac{Positivos\ Verdadeiros + Negativos\ Verdadeiros}{Positivos\ Verdadeiros + Negativos\ Verdadeiros + Positivos\ Falsos + Negativos\ Falsos}$;
B. precisão ("precision"), a confiabilidade do modelo quando ele aponta que um resultado é positivo: $\frac{Positivos\ Verdadeiros}{Positivos\ Verdadeiros + Positivos\ Falsos}$;
C. revocação ("recall"), a confiabilidade do modelo em detectar *todos* os resultados positivos corretamente: $\frac{Positivos\ Verdadeiros}{Positivos\ Verdadeiros + Negativos\ Falsos}$;

4.1.4. Value Proposition Canvas

Perfil do Cliente



4.1.5. Matriz de Riscos

		Ameaças				
Probabilidade	90%			Bugs no código do modelo preditivo		
	70%					
	50%			Imprecisão / inconsistência dos dados usados para análise		
	30%			Não ter dados relevantes para solucionar o problema	Modelo ter uma precisão ou acurácia baixa	
	10%					Sistema exigir muita capacidade computacional
		Muito baixo	Baixo	Moderado	Alto	Muito alto
		Impacto				

		Oportunidades				
Probabilidade	90%					
	70%	Concorrência não possuir ferramenta que mapeie o churn	Sistema prevê corretamente a chance de saída dos RTs	Bom volume de dados para treinar o modelo		
	50%					
	30%					
	10%					
		Muito alto	Alto	Moderado	Baixo	Muito baixo
		Impacto				

4.1.6. Personas

Posicione aqui suas Personas (as que utilizam o modelo e as que são afetadas pelo modelo)

4.1.7. Jornadas do Usuário

4.2. Compreensão dos Dados

4.2.1. Descrição Geral dos Dados

Os dados disponibilizados pelo cliente consistem em diversos arquivos "Comma-separated values" ("CSV"), elencados na tabela abaixo. Segundo informações repassadas pela Rappi, todos os dados foram obtidos através do aplicativo. Também descrevemos na tabela o que cada arquivo representa, em geral, e o seu tamanho, com número de linhas e de colunas.

Nome do CSV	Descrição Geral do Conteúdo	Número de Linhas	Número de Colunas
attendance_rate	Taxa de aceitação dos pedidos. Ex: tocou 10 vezes, aceitei 9, 90%	653.167	2
comp defects	Informações sobre pedido com algo incompleto, faltante, item errado, etc, (literalmente qualquer coisa que acontecer). Caso seja falha Rappi/RT o cliente é reembolsado. Esse arquivo contém um registro desses reembolsos.	6.783.958	10
criacao contas churn	Data de criação das contas churn do período	32.568.384	9
earnings	Receita de cada entregador	566.099	4
Incidentes_Regras RT	Existem diversas regras para melhoria da qualidade da operação. Exemplo: A regra 92/93 remove/libera o RT do pedido caso ele não esteja em movimento ou em direção ao cliente. (Reforço que não temos informação sobre a localização do RT, apenas se esta	2.405.601	9

	diminuindo o ETA ou não).		
Infos Gerais	Informações gerais do entregador, inclusive o último pedido ("churn" é um entregador com último pedido há mais de 21 dias)	180.178	25
Ordens Done e Cancel	ordens realizadas e ordens canceladas; Ordens podem ser canceladas por qualquer razão: pela loja, falta de produto, pelo RT (Pneu furado, roubo, acidente, problemas pessoais, etc).	653.166	4
Product return	Retorno de produto uma vez que a ordem foi cancelada. Ex: comprei itens de supermercado e por qualquer razão ordem foi cancelada; o RT precisa retornar este a loja (nem todas aceitam, ponto ruim) ou devolver em algum ponto de apoio Rappi; Até isso acontecer, ele fica com uma dívida no valor dos produtos;	41.535	11
supply	Tempo em Horas que o RT fica/ficou conectado no período;	124.526	10
Tempo de Resolução e Modal	Quanto tempo o entregador ficou esperando resolução para algum pedido aberto junto ao suporte Central Rappi. *OBSERVAÇÃO IMPORTANTE: o arquivo repassado é igual ao CSV "criação contas churn", sem informações sobre tempo de resolução. Acreditamos que esse não é o arquivo correto.*	32.568.384	9

Mais especificamente, podemos destacar o que cada coluna significa nos arquivos CSV, conforme pode ser visto na tabela abaixo:

CSV	Nome da Coluna	Descrição	Tipo	Dúvidas
attendance_rate	STOREKEEPER_ID	ID do entregador	float64	
attendance_rate	ACCEPTANCE_RATE	Porcentagem de pedidos aceitos pelo entregador em relação ao total de pedidos recebidos	float64	R: Esse indicador influencia no nível do entregador. Também pode sofrer medidas disciplinares (bloqueio, bloqueios temporários). Muita gente está em block temporário. A maior parte fala que são bloqueios injustos.
comp_defects	STOREKEEPER_ID	ID do entregador	float64	
comp_defects	WEEK	Segunda-feira da semana em referência (YYYY-MM-DD)	object	
comp_defects	CITY	Cidade	object	
comp_defects	LEVEL_ID	ID do nível do entregador	float64	
comp_defects	LEVEL_NAME	Nome do nível do entregador	object	P: Quais são os níveis e as classificações do level_name? R: Diamante (melhor), Prata, Bronze e Danger (pior, muito propenso a sair).
comp_defects	ORDERS	Número de pedidos que aconteceu algum problema	int64	
comp_defects	GMV_TOTAL	Total da transação (GMV = custo total pago pela Rappi para a loja)	float64	

comp_defects	COMPENSATIONS	Valor pago (devolvido) para o usuário	float64	
comp_defects	DEFECT_COMPENSATIONS	São as ordens.	float64	P: Esse valor é uma quantia ou ID? R: ID do pedido.
comp_defects	DEFECT_ORDER	São as ordens.	float64	P: Esse valor é uma quantia ou ID? R: ID do pedido.
criacao contas churn	ID	ID do entregador	int64	
criacao contas churn	FIRST_NAME	Nome do entregador	object	
criacao contas churn	GENDER	Gênero do entregador	object	
criacao contas churn	CITY	Cidade de atuação	object	
criacao contas churn	SK.CREATED_AT::DATE	Data de cadastro da conta	object	
criacao contas churn	TRANSPORT_MEDIA_TYPE	Modal de transporte	object	
criacao contas churn	CARTAO	Tem cartão pré-pago?	object	
criacao contas churn	LEVEL_NAME	Nível do entregador	object	
criacao contas churn	FECHA_ULT	Última data em que o entregador interagiu no aplicativo	object	P: Essa é a última data de entrada no aplicativo ou a data da última entrega? R: Última data em que houve interação com o aplicativo.
earnings	STOREKEEPER_ID	ID do comerciante	int64	

earnings	MONTH	Primeiro dia do Mês, em formato ISO-8601	object	Receita que o entregador ganhou no mês em referência
earnings	EARNINGS	Receita do entregador	float64	P: A receita está em reais? É realmente esse valor ou devemos multiplicar por algum múltiplo (10X, 100X)? R: Ela está em dólares (taxa de conversão: 4.45). É o líquido recebido pelo entregador.
earnings	TIPS	Gorjetas do entregador	float64	R: Tips são gorjetas recebidas pelos entregadores. 100% do entregador
Incidentes_Regras RT	DATE	Data do Incidente (formato YY-MM-DD)	object	
Incidentes_Regras RT	NAME	Nome do Incidente	object	P: Qual o significado desses códigos (exemplo: 92. Liberación (Live), 75. Reporte Manual). R: Essa informação é muito sensível, não pode ser compartilhada. Por exemplo, um deles muito comum é o entregador ficar parado / aumentar distância e não entregar o pedido.
Incidentes_Regras RT	INCIDENT_ID	ID do incidente	int64	
Incidentes_Regras RT	STOREKEEPER_ID	ID do entregador	int64	
Incidentes_Regras RT	PUNISHMENT_MINUTES	Minutos de suspensão	int64	
Incidentes_Regras RT	PUNISHMENT_TYPE	Tipo de punição (permanent_block, temporary_block, warning)	object	R: Warning: É um aviso no aplicativo. A punição depende do nível do RT. Por ex. Warning para Diamante e 15 min de bloqueio para os demais. Isso inclusive pode motivar o "churn".

Incidentes_Regras RT	DISCIPLINE_RULE_BUCKET	É uma derivada de um indicador interno da Rappi	object	
Incidentes_Regras RT	CATEGORY_RULE	Gênero da regra aplicável	object	Qual a diferença entre "Discipline" e "Manual"? R: Discipline seria uma regra "educativa" automática, para ajudar o RT. Manual é um bloqueio temporário ou permanente feito por um ser humano (suporte da Rappi).
Incidentes_Regras RT	ORDER_ID	ID do pedido	float64	
Infos Gerais	ID	ID do entregador	int64	
Infos Gerais	NOME	Nome	object	
Infos Gerais	SOBRENOME	Sobrenome	object	
Infos Gerais	GENERO	Gênero	object	
Infos Gerais	DATA_NASCIMENTO	Data de Nascimento	object	
Infos Gerais	CIDADE	Cidade de atuação	object	
Infos Gerais	IS_ACTIVE	Se o entregador está ativo ou não	bool	
Infos Gerais	TRANSPORTE	Modal de transporte	object	
Infos Gerais	AUTO_ACEITE	Se o entregador aceita pedidos automaticamente	bool	O auto aceite é uma opção no aplicativo.
Infos Gerais	COUNT_ORDERS_LAST_7D	Ordens realizadas nos últimos 7 dias	int64	

Infos Gerais	COUNT_ORDERS_LAST_30D	Ordens realizadas nos últimos 30 dias	int64	
Infos Gerais	COUNT_ORDERS_CANCELED_LAST_7D	Ordens canceladas nos últimos 7 dias	int64	
Infos Gerais	COUNT_ORDERS_CANCELED_LAST_30D	Ordens canceladas nos últimos 30 dias	int64	
Infos Gerais	GORJETA	Gorjetas recebidas pelo entregado	float64	
Infos Gerais	PRIMEIRO_PEDIDO	Data do primeiro pedido entregue	object	
Infos Gerais	ULTIMO_PEDIDO	Data do último pedido entregue	object	
Infos Gerais	COUNT_ORDERS_RESTAURANTES	Ordens de restaurantes	int64	
Infos Gerais	COUNT_ORDERS_MERCADO	Ordens de mercado	int64	
Infos Gerais	COUNT_ORDERS_FARMACIA	Ordens de farmácia	int64	
Infos Gerais	COUNT_ORDERS_EXPRESS	Ordens turbo (10min)	int64	
Infos Gerais	COUNT_ORDERS_ECOMMERCE	Ordens e-commerce	int64	
Infos Gerais	COUNT_ORDERS_ANTOJO	Ordens loja fora da plataforma	int64	

Infos Gerais	FRETE_MEDIO	Média dos fretes recebidos nas entregas	float64	
Infos Gerais	COOKING_TIME_MEDIO	Média de tempo esperando o pedido ficar pronto nos restaurantes	float64	
Infos Gerais	ITENS_MEDIO	Média de itens nas entregas	float64	
Ordens Done e Cancel	STOREKEEPER_ID	ID do comerciante	int64	
Ordens Done e Cancel	ORDERS_DONE	Número de pedidos realizados	int64	P: Esse número é o total de pedidos atendidos na plataforma ou existe um limite temporal (mês, ano)? R: É o total da vida toda.
Ordens Done e Cancel	ORDERS_CANCEL	Ordens totais canceladas pelo entregador	int64	P: Se um pedido é cancelado, ele é contabilizado só em ORDERS_CANCEL ou ele é primeiro contabilizado em ORDERS_DONE quando realizado e depois também em ORDERS_CANCEL quando cancelado? R: Se uma ordem é cancelada ela entra apenas em canceladas.
Ordens Done e Cancel	CANCEL_OPS_RT	Ordens canceladas manualmente pelo time de operação	int64	
Product return	ID_ENTREGADOR	ID do entregador	int64	Esse ID está fora do padrão "STOREKEEPER_ID" de entregador (em português, com ID antes do nome), mas são a mesma coisa.

Product return	LEVEL_NAME	Nível do entregador	object	
Product return	MODAL	Meio de locomoção	object	
Product return	CITY	Cidade de atuação	object	
Product return	CREATED_AT	Data de cadastro	object	
Product return	ORDER_ID	ID do pedido	int64	
Product return	PRODUCT_RETURNS	Valor da devolução (negativo)	float64	P: Por ser negativo, esse valor é subtraído de algum outro? R: Sim, do GMV, caso o COUNT_TO_GMV seja positivo.
Product return	VERTICAL_SUB_GROUP	Categoria do pedido (farmácia, mercado)	object	
Product return	COUNT_TO_GMV	Abater do GMV	bool	P: Esse booleano significa que a devolução deve ser abatida do GMV? R: Alguns pedidos não contam para o GMV. Se estiver falso ele não está sendo contabilizado no resultado.
Product return	GMV	Gross Merchandise Value	float64	GMV significa Gross Merchandise Value (vendas * mercadorias vendidas). Isso é o valor do pedido? R: É o valor total pago no pedido pela Rappi.
Product return	STORE_ID	ID da loja	int64	
supply	STOREKEEPER_ID	ID do entregador	int64	
supply	CITY	Cidade de atuação	object	

supply	DATE	Dia da semana (YYYY-MM-DD)	object	R: Essa é a data que conta.
supply	WEEK	Segunda-feira da semana em referência (YYYY-MM-DD)	object	
supply	CREATED_CARD	Data de Criação do Cartão	object	
supply	LEVEL_NAME_2	Nível do entregador	object	P: Por que esse campo "level_name" tem um "_2" no final? Quais são os níveis e as classificações desse "level_name"? R: É a mesma coisa que "level_name". Igual ao "ID_ENTREGADOR", é uma inconsistência do banco de dados.
supply	HAVE_CARD	Tem um cartão?	bool	P: O que é esse cartão? R: É um cartão pré-pago que o entregador pode ter para fazer as compras. O dinheiro é liberado na hora do pagamento. Também existe a integração "Cashless", que é uma integração da Rappi com a loja.
supply	TRANSPORT_MEDIA_TYPE	Modal de transporte	object	
supply	NUM_ORDERS	Número de pedidos atendidos pelo entregador naquele dia	int64	
supply	SUPPLY_HOURS	Horas em que o entregador ficou conectado no período	float64	

Forma de agregação e mesclagem de dados:

Todos os arquivos CSV tem uma coluna com um ID único do entregador (geralmente a coluna tem o nome "STOREKEEPER_ID", mas em alguns CSVs pode ter um nome diferente como "ID_ENTREGADOR"). Desse modo, podemos usar esse identificador único para juntar as diferentes tabelas (fazer um "join") e assim cruzar as informações de um CSV com o outro.

Riscos e contingências relacionados aos dados (qualidade, cobertura/diversidade e acesso):

Em relação a qualidade, temos o risco que os arquivos repassados possam ter dados que não são verossímeis por uma série de motivos (inserções errôneas, importações indevidas, "bugs" no aplicativo ou base de dados). Adicionalmente, podemos verificar um nível elevado de ruído em alguns arquivos CSV, como informações ausentes (campos nulos) e muitos outliers (por exemplo, muitos entregadores com receitas além do esperado no arquivo "earnings").

Sobre cobertura e diversidade, temos alguns arquivos com muitas informações, como o "criacao contas churn" com mais de 32 milhões de registros. Isso pode levar a um excesso de informações para tratamento e análise, caso não seja feito um cruzamento e filtragem prévios.

Finalmente, no que concerne ao acesso, a Rappi não pode liberar alguns dados, como as regras automáticas usadas para aplicar suspensões e avisos na plataforma (por exemplo, quanto tempo um entregador pode ficar parado até receber um aviso). Assim, isso pode dificultar algumas análises.

Seleção do subconjunto para análises iniciais:

Temos algumas hipóteses iniciais sobre as causas do "churn", que irão embasar nossas análises iniciais. Identificamos essas hipóteses a seguir, junto com o conjunto de dados relevante:

- *Hipótese 1: Os entregadores ("RT"s) deixam a plataforma por muitas suspensões recorrentes.*
    - Subconjunto relevante: arquivo "Incidentes_Regras RT", com as suspensões. Verificar se existe uma relação entre número e duração de suspensões e "churn".

- *Hipótese 2: Os entregadores deixam a plataforma por devoluções excessivas de pedidos, que levam a dívidas no valor dos produtos e custos com devoluções (gasolina, tempo de deslocamento).*
    - Subconjunto relevante: arquivo "comp defects", com as devoluções pendentes. Verificar se existe uma relação entre devoluções e "churn".

- *Hipótese 3: Os entregadores deixam a plataforma por baixa remuneração ou poucos pedidos.*
    - Subconjunto relevante: arquivos "earnings", "Ordens done e cancel" e, especialmente, "Infos gerais". Verificar se existe uma relação entre rendimentos, número de ordens realizadas e "churn".

Restrições de segurança: foi mencionado que são dados sensíveis e que devem ser geridos com cuidado. Portanto, não vamos publicar essa base de dados, nem repassar essas informações para terceiros. O uso dos dados será exclusivamente para desenvolver um modelo preditivo para a possibilidade de "churn" de entregadores na plataforma Rappi.

4.2.2. Estatística Descritiva dos Dados

A partir dos dados fornecidos, podemos estabelecer algumas análises pautadas em estatística descritiva e elaborar gráficos para testar as hipóteses suscitadas inicialmente.

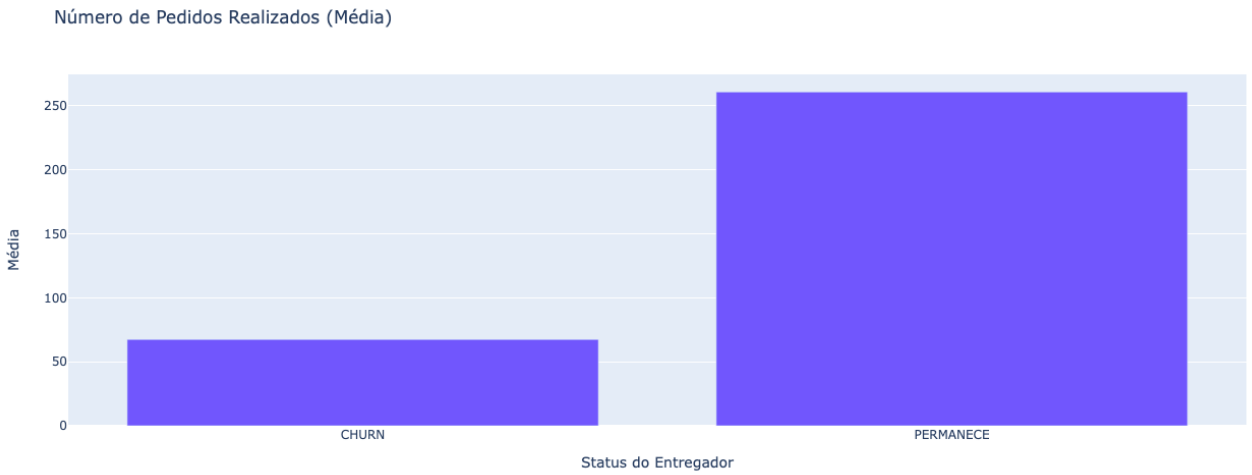
- *Hipótese A: Os entregadores deixam a plataforma por baixa remuneração ou poucos pedidos*

Para fins de comparação para essa hipótese, foi necessário trabalhar com dois grupos – os entregadores que saíram da plataforma e aqueles que permanecem na mesma. Para tanto, dividimos o arquivo "Infos Gerais" em dois grupos, os que entregaram o último pedido há mais de 21 dias ("CHURN")- critério informado em entrevista com a Rappi para classificação de Churn - e os que entregaram pedidos após essa data ("PERMANECE"):

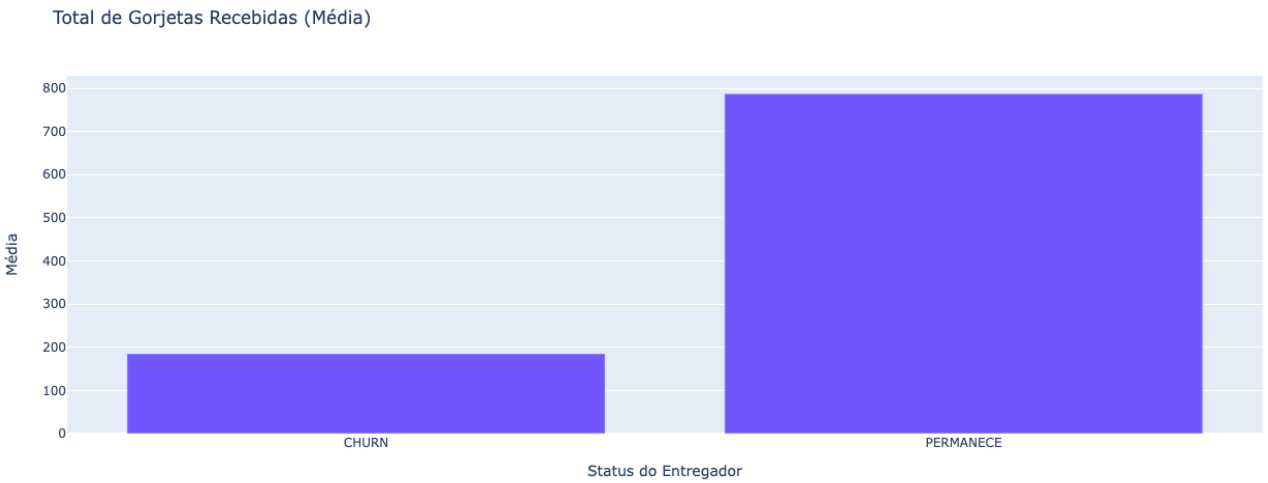
```
df_infos = pd.read_csv('/content/drive/MyDrive/data/infos gerais.csv')
df_infos.sort_values('ULTIMO_PEDIDO', ascending=False)
# Churns são entregadores com o último pedido há mais de 21 dias
# Os dados vão até 01/08/2022, logo a data de corte é 11/07/2022
df_not_churned = df_infos[(df_infos['ULTIMO_PEDIDO'] > '2022-07-11')]
df_churned = df_infos[(df_infos['ULTIMO_PEDIDO'] < '2022-07-11')]
```

Realizando uma comparação entre os dois grupos, obtivemos resultados interessantes.

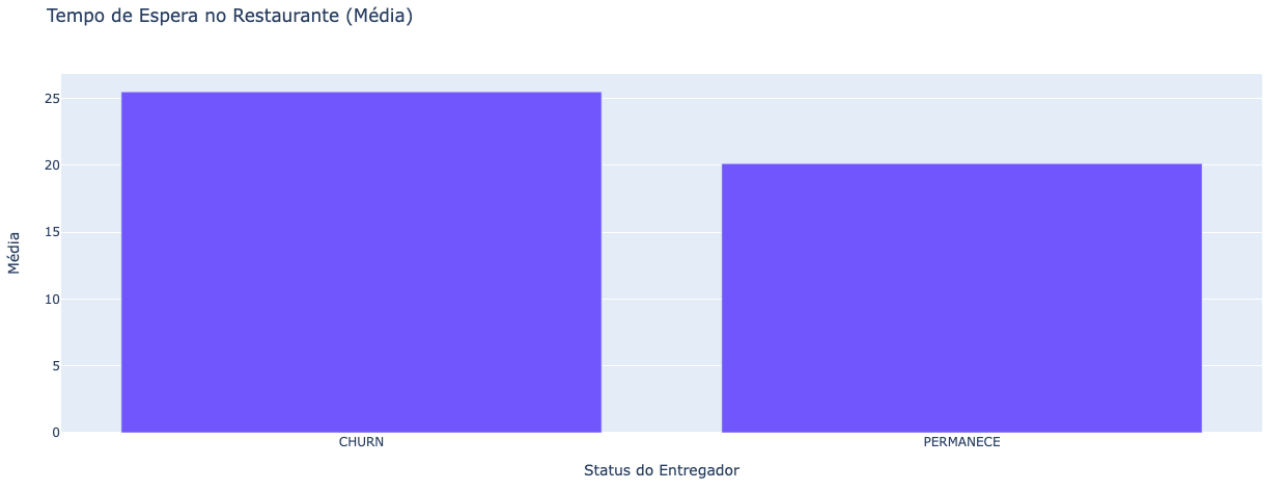
- Os entregadores que sofreram "churn" realizaram menos entregas do que aqueles que permaneceram na plataforma, sendo o número de pedidos cerca de 4 vezes maior. Isso pode ser um indicador de que quanto mais corridas o entregador recebe, menor sua propensão a sair da plataforma:



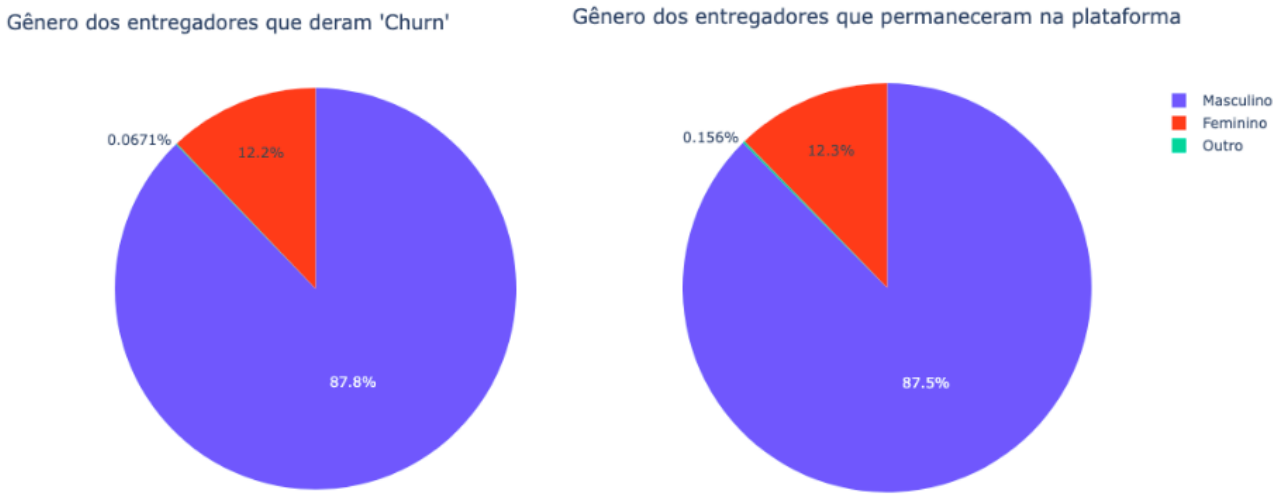
- Os entregadores que permaneceram na plataforma ganharam quase o quádruplo de gorjetas daqueles que deram "churn". Isso pode indicar que as gorjetas (ou ausência delas) são um fator de permanência na plataforma:



- Entregadores que permaneceram na plataforma em geral esperam menos tempo para o pedido ficar pronto no restaurante. Não conseguimos entender qual a causa desse fator:

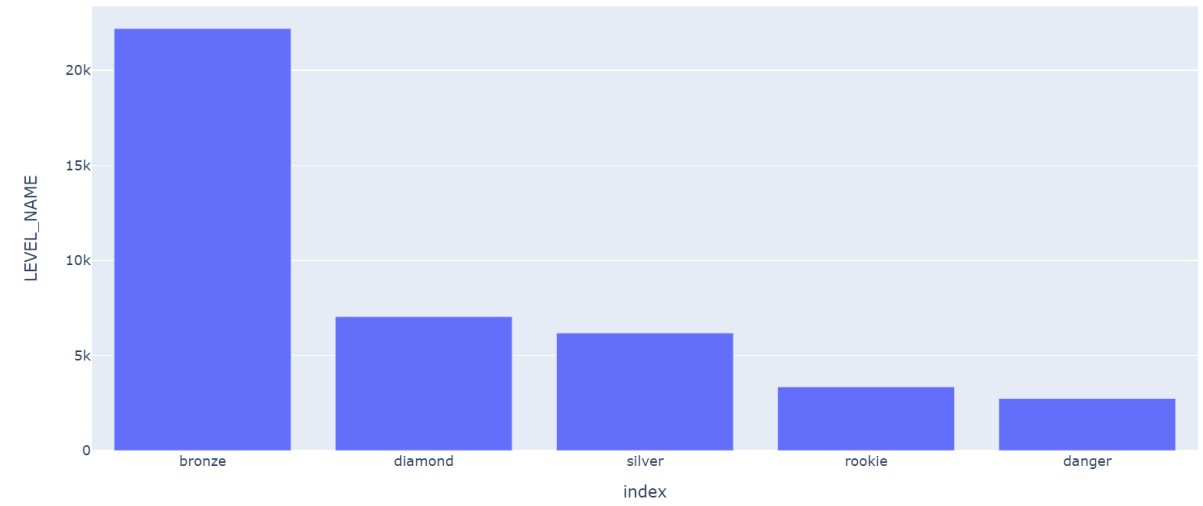


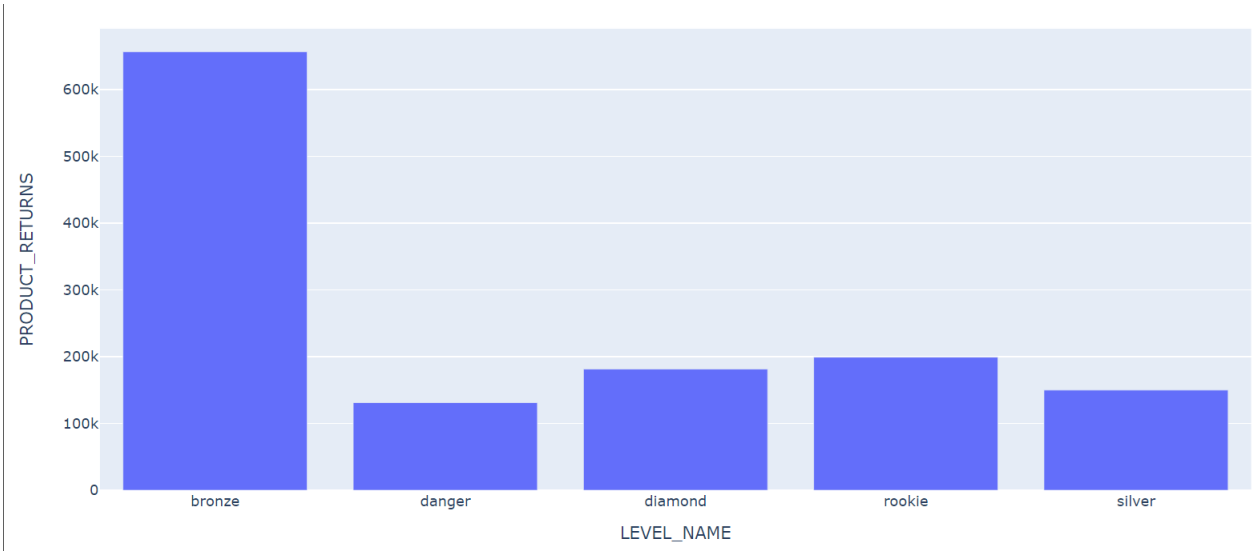
- Aspectos demográficos, como gênero do entregador, aparentam não impactar de forma expressiva o "churn". A composição dos dois grupos é praticamente idêntica:



- *Hipótese B: Os entregadores deixam a plataforma por devoluções excessivas de pedidos*

    - Para trabalhar com essa hipótese, verificamos se algum grupo está representado de forma desproporcional no valor de pedidos a devolver pendentes. Assim, comparamos a quantidade de entregadores em cada nível e a quantidade de devoluções pendentes:





Nesse contexto, verificamos que há uma quantidade expressiva de devolução no nível bronze (porém proporcional), e uma quantidade muito acima do esperado no nível "rookie" e "danger". O que mais se destaca nessa análise é que o nível "bronze" pode representar um grupo descontente com a plataforma, dado o alto volume de devoluções pendentes.

4.2.3. Descrição da Predição Desejada ("Target")

A predição desejada ("target") consiste em classificar se um entregador ("RT") irá sair da plataforma Rappi (evento de "churn") ou se irá permanecer na plataforma.

Diferentemente de um modelo de regressão contínua, em que se deseja encontrar um valor numérico específico por interpolação (por exemplo, o valor de uma casa em determinado mercado imobiliário), no modelo de classificação o mais importante é prever em qual categoria se deve encaixar uma situação (por exemplo, um paciente tem uma patologia ou não tem).

Nosso caso, portanto, consiste em um modelo de **classificação binária**, em que o "target" pode ser encaixado em uma de duas categorias: a) o entregador saiu; *ou* b) permaneceu na plataforma durante determinado período.

4.3. Preparação dos Dados

Descreva as etapas realizadas para definir os dados e os atributos descritivos dos dados (“features”) a serem utilizados. Essa descrição deve ser feita de modo a garantir uma futura reprodução do processo por outras pessoas, e deve conter:

a) Descrição de quaisquer manipulações necessárias nos registros e suas respectivas features.

b) Se aplicável, como deve ser feita a agregação de registros e/ou derivação de novos atributos.

c) Se aplicável, como devem ser removidos ou substituídos valores ausentes/em branco.

d) Identificação das features selecionadas, com descrição dos motivos de seleção.

Não deixe de usar tabelas e gráficos de visualização de dados para melhor ilustrar suas descrições.

IMPORTANTE: Crie tópicos utilizando a formatação “Heading 3” (ou menor) para que o Google Docs identifique e atualize o Sumário (é necessário apertar o botão Refresh no Sumário para ele coletar as atualizações)

4.4. Modelagem

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

4.5. Avaliação

Nesta seção, descreva a solução final de modelo preditivo, e justifique a escolha. Alinhe sua justificativa com a seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

4.6 Comparação de Modelos

5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

6. Referências

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.