Documentação Modelo Preditivo - Inteli

Forecast

Nome do grupo: ForeSee

Integrantes: Davi Rosalino Glória Motta, Diogo Pelaes Burgierman, João Pedro Brandão de Moura, Lucas Nogueira Storelli de Luccas, Raí de Oliveira Cajé e Renan Feitosa Oliveira

Sumário

1. Introdução

O seguinte projeto tem como objetivo auxiliar a PowerCo, uma das maiores operadoras de redes de energia e infraestrutura de energia da Europa e um fornecedor de soluções inovadoras para aproximadamente 48 milhões de clientes.

Segundo o website da empresa, os valores da PowerCo são fundamentados em um modelo de negócios voltado para o cliente, exploração de novas oportunidades, produtos e serviços e responsabilidade de entregas que atinjam o nível das expectativas criadas.

A PowerCo acredita que somente de acordo com esses valores pode cumprir a missão de ser uma das maiores operadoras de redes e infraestrutura de energia da Europa e fornecedora de soluções para o vasto número de clientes, impulsionando de forma decisiva a transição energética na Europa e usando com sucesso os negócios para promover a sustentabilidade e a proteção do clima.

Nesse contexto, dada a ênfase que a PowerCo coloca em satisfazer as necessidades de seus clientes, é de grande importância que ela seja capaz de manter e suprir as expectativas dos clientes para que suas estratégias globais se concretizem com sucesso. No entanto, a empresa enfrenta desafios no que diz respeito às taxas de rotatividade, o que não só contrapõe o modelo de negócios estabelecido, mas também resulta em custos significativos associados à aquisição de novos clientes.

2. Objetivos e Justificativa

2.1 Objetivos

O propósito central deste projeto é elaborar um modelo preditivo por meio da exploração do banco de dados de clientes da PowerCo, atuando diretamente na identificação de padrões na rotatividade de clientes. Nesse contexto, o algoritmo terá a capacidade de analisar informações e prever quais clientes têm maior probabilidade de se desligarem da empresa.

Em seguida, serão fornecidos ao usuário métodos para mitigar essa perda, considerando como principal abordagem a aplicação de descontos mínimos. Essa abordagem visa manter a atratividade dos produtos/serviços, mantendo os clientes satisfeitos com a relação custo-benefício. Isso não só se alinha aos valores da PowerCo, mas também otimiza os custos de aquisição de novos clientes.

Além disso, o sistema fornecerá resultados simplificados, tendo como objetivo uma fácil compreensão por partes interessadas na análise da rotatividade de clientes, mesmo que não possuam conhecimento prévio na área de ciência de dados.

2.2 Proposta de solução

Através da análise abrangente dos dados de clientes da PowerCo, o modelo se concentrará em identificar padrões e tendências que indicam a probabilidade de um cliente optar pelo desligamento. Isso permitirá que a PowerCo antecipe e tome medidas preventivas.

Para cumprir nossos objetivos, o modelo utilizará algoritmos de aprendizado de máquina avançados, treinados com dados históricos e variáveis relevantes, como padrões de consumo, durações de contrato, estimativas de preços futuros e histórico de desligamentos. O resultado será uma pontuação de probabilidade de desligamento para cada cliente.

Para a implementação desse projeto, adotaremos o ambiente do Google Colab e a linguagem de programação Python será a base fundamental. Ademais, contaremos com bibliotecas de análise de dados robustas, tais como Pandas, Seaborn e Matplotlib, que potencializam nossa capacidade de explorar e visualizar os dados, bem como identificar padrões cruciais.

2.3 Justificativa

Em primeiro lugar, a abordagem deste projeto utiliza várias informações históricas e comportamentais dos clientes para criar previsões de desligamento precisas. Isso permite à PowerCo antecipar problemas e agir proativamente para reter clientes em risco, em vez de simplesmente reagir após o desligamento ter ocorrido. Essa capacidade preditiva pode economizar recursos significativos ao evitar os custos associados à aquisição de novos clientes.

Além disso, a estratégia de oferecer descontos mínimos personalizados se destaca como uma maneira inovadora de retenção de clientes. Ao equilibrar cuidadosamente o valor oferecido aos clientes com a viabilidade financeira da PowerCo, essa abordagem aumenta as chances de manter clientes enquanto otimiza o impacto nos resultados financeiros.

Por último, o uso das bibliotecas de análise de dados, como Pandas, Seaborn e Matplotlib, amplia nossa capacidade de explorar e visualizar os dados, permitindo uma análise mais profunda e uma comunicação eficaz dos insights. Isso se diferencia como uma abordagem prática, que não apenas fornece resultados, mas também os torna compreensíveis e acionáveis para diversas partes interessadas.

3. Metodologia

Ao trabalhar com Machine Learning e análise de dados, é necessário saber trabalhar com o dataset de uma forma funcional, traçando planos de ações e estratégias para tornar o trabalho mais assertivo. Para isso, foi utilizada a metodologia CRISP DM durante o desenvolvimento do projeto.

Segundo o grupo Voitto, CRISP DM ou Cross Industry Standard Process for Data Mining (Processo Padrão Inter-Indústrias para Mineração de Dados) é uma metodologia ágil que fornece uma abordagem estruturada e robusta para o planejamento de projetos envolvendo Machine Learning, mineração e análise de dados.

Para o projeto realizado, a metodologia foi de grande importância. Por operar como um ciclo iterativo, permite regressões às fases anteriores quando necessário. Assim, traz flexibilidade e dinamismo durante o trabalho. Por exemplo, se durante a modelagem de dados o grupo perceber uma defasagem em alguma coluna, pode retornar para a etapa de preparação e traçar novos planos e objetivos.

É formado por **6 etapas**:

Entendendo o negócio:

Neste estágio, procederemos com uma análise aprofundada do projeto ou empreendimento (compreensão do negócio), em conformidade com as metas e interesses do cliente. Será essencial identificar eventuais obstáculos e fatores que possam afetar o resultado final do projeto. Nessa fase, é crucial estabelecer de forma precisa os objetivos, metas, potenciais barreiras, e riscos associados, bem como as aplicações previstas para o produto a ser desenvolvido. Também é importante abordar questões relacionadas aos custos, terminologia e os critérios de êxito no contexto empresarial. Além disso, é fundamental avaliar os recursos disponíveis na empresa, incluindo ferramentas, software, banco de dados, e outros ativos relevantes. A partir dessas análises, poderemos iniciar o planejamento das ações necessárias.

Entendendo os dados:

Nesta fase, é conduzida uma análise abrangente do projeto ou empreendimento, alinhando-o com os objetivos e interesses do cliente. É necessário durante essa etapa

identificar potenciais obstáculos e variáveis que possam influenciar o desfecho final do projeto.

Neste estágio crucial, é fundamental estabelecer de maneira transparente os seguintes elementos: metas, objetivos, potenciais entraves, riscos, aplicações para o produto a ser desenvolvido, custos associados, terminologia essencial e os critérios que definirão o sucesso empresarial.

Também é de suma importância avaliar os recursos à disposição da empresa, incluindo ferramentas, software, bancos de dados e outros ativos relevantes. É a partir dessas informações que começaremos a elaborar os planos de ação necessários para dar início ao projeto.

Preparando os dados:

A preparação de dados desempenha um papel fundamental na qualidade dos resultados de análises e modelagens de dados. Ela começa com a seleção cuidadosa dos dados a serem utilizados, pois dados de entrada inadequados podem levar a resultados imprecisos. Em seguida, a limpeza de dados entra em ação, onde verificamos a presença de dados corrompidos ou inconsistentes e os eliminamos do conjunto de dados. Isso é essencial para manter a integridade dos dados. A construção de dados é outro aspecto importante, envolvendo a criação de novos conjuntos de dados derivados dos dados originais. Isso pode incluir a geração de novas variáveis ou recursos que podem ser mais informativos para o problema em questão. A integração de dados é um passo crucial, no qual diferentes fontes de dados são unidas ou mescladas para criar um conjunto de dados mais abrangente e consistente. Isso pode melhorar a qualidade dos dados e enriquecer as informações disponíveis para a etapa de modelagem.

Modelagem:

Modelagem (Modeling) envolve a aplicação de técnicas e algoritmos específicos para criar modelos preditivos ou descritivos com base nos dados. Embora muitos algoritmos de Machine Learning, como árvores de decisão, redes neurais e regressão logística, possam ser usados para tarefas de classificação, eles também podem ser aplicados a outras tarefas, como regressão, clustering, entre outros. Neste estágio, é comum que se faça um loop contínuo com o estágio de Preparação dos Dados. Isso ocorre porque, ao tentar modelar os dados, podemos descobrir a necessidade de mais limpeza, transformação ou engenharia de recursos. É essencial dividir os dados em pelo menos dois conjuntos: treino e teste. O conjunto de treino é utilizado para construir e treinar o modelo, enquanto o conjunto de teste serve para avaliar a performance do modelo em dados não vistos anteriormente. Em muitos casos, também se utiliza um terceiro conjunto chamado de validação, que auxilia na otimização dos hiperparâmetros do modelo. Durante a fase de Modelagem, a equipe selecionará o algoritmo mais adequado ao problema, definirá planos de testes para validação, construirá o modelo e, por fim, avaliará sua performance e adequação aos objetivos do projeto.

Avaliação:

A etapa de Avaliação (Evaluation) tem como principal objetivo avaliar a qualidade e a fidedignidade do modelo gerado na etapa de Modelagem. Durante essa fase, é fundamental revisitar os objetivos estabelecidos no estágio inicial de Definição de Objetivos para garantir que os resultados do modelo estejam alinhados com as metas e expectativas do projeto. A avaliação não se restringe apenas à performance do modelo, mas também à sua aplicabilidade prática e à segurança dos resultados gerados. É comum, ao longo dessa etapa, identificar novas necessidades ou padrões de dados não previstos anteriormente. Esse reconhecimento pode implicar em ajustes no modelo ou até mesmo na coleta e preparação dos dados. Em algumas situações, pode ser necessário retornar a etapas anteriores do CRISP-DM para refinar o trabalho realizado. Ao final da etapa de Avaliação, as decisões sobre os próximos passos do projeto são tomadas, sejam elas a implementação do modelo, sua revisão ou até mesmo a redefinição de objetivos. É essencial que essa etapa forneça clareza sobre as ações a serem seguidas, garantindo que o modelo possa ser utilizado com confiança e eficácia.

Implantação:

Na fase de implantação, o processo de desenvolvimento dos modelos previamente criados e avaliados nas etapas anteriores é iniciado. É importante destacar que essa etapa somente é viável quando todos os objetivos das fases anteriores foram alcançados com êxito. Neste estágio, a ênfase está na integração dos modelos em ambientes de produção. A implantação pode ocorrer por meio de pipelines de implementação ou através de serviços de computação em nuvem. Os principais objetivos a serem alcançados durante essa etapa são os seguintes:

- Planejamento da implantação: Consiste na efetiva implantação do software e dos modelos em produção, garantindo que todos os elementos estejam funcionando corretamente.

- Monitoramento e manutenção: É essencial acompanhar continuamente o desempenho dos modelos em ambiente real, fazendo ajustes e correções conforme necessário para garantir que eles continuem a atender aos requisitos.

- Geração de relatórios: Documentar de maneira abrangente todos os processos e resultados é fundamental para manter um registro claro do desempenho e dos resultados obtidos.

- Avaliação dos resultados finais; Aferir o impacto dos modelos implantados em diversos aspectos, como financeiros, comerciais, de marketing, recursos humanos, produção, entre outros, é crucial para compreender o impacto e a eficácia do modelo no ambiente operacional. Esta etapa analisa o desempenho quantitativo, como ROI e métricas tradicionais de Machine Learning, e o impacto qualitativo, incluindo benefícios intangíveis e satisfação do cliente. É essencial comparar os resultados com os objetivos iniciais, coletar feedback dos usuários e entender o contexto mais amplo de operação do modelo. Concluindo, esta avaliação não só valida a eficácia do modelo, mas também reconhece seu valor no panorama geral do negócio, estabelecendo bases para futuras melhorias.

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

A PowerCo está situada no mercado de energia europeu como uma empresa de fornecimento de gás e eletricidade. Isto significa que, apesar de estar localizada em um continente com grande mercado consumidor de energia, ela é diretamente afetada por eventos como a guerra entre Rússia e Ucrânia, bem como a crescente busca por fontes de energia limpas e renováveis.

De forma sucinta, em decorrência da Guerra Russo-ucraniana, grande parte do fornecimento de energia e gás da Rússia para a Europa foi cortado. A União Europeia, por sua vez, buscou contornar a crise energética advinda desta situação por meio do incentivo ao uso de fontes de energia limpas e renováveis para aumentar a autonomia energética do bloco.

Com isso, apesar das empresas europeias de fornecimento de energia fóssil terem experienciado (pouco tempo após os cortes de fornecimento de energia russa) um cenário de mercado favorável, estas enfrentam uma competitividade significativa com companhias que fornecem energia renovável e/ou limpa. Esse cenário, acrescido da constante busca pelos melhores preços e descontos pelos consumidores de energia, configura o contexto sociopolítico e econômico no qual a PowerCo se insere.

Desta forma, é interessante analisar o cenário de mercado da PowerCo de forma mais profunda. Para isso, foi feita a análise das “5 Forças de Porter”, Framework apresentado por Michael Porter em seu livro ‘Estratégia Competitiva’ (1980), tal análise é um modelo para definir e descrever a intensidade e o modo de cada um dos fatores externos que influenciam na presença de mercado de uma empresa. São estes a *rivalidade entre concorrentes*, *poder de barganha dos fornecedores*, *poder de barganha dos compradores*, *ameaça de novos entrantes*, *ameaça de produtos ou serviços substitutos*, os quais são descritos a seguir:

Quadro 1 - 5 Forças de Porter

Fonte: Autoria própria

Produtos Substitutos: Os substitutos, segundo Michael Porter, são os produtos capazes de inclinar o consumidor a trocar aqueles oferecidos pela empresa em questão por outros que apresentem maior rendimento, menores custos, ou outras características vantajosas. Assim sendo, dado que a matriz energética da PowerCo é baseada em fontes não renováveis e, na atualidade, é notório um apelo maior à consciência ambiental e sustentabilidade, vale citar, como produto substituto aos oferecidos pela empresa, a energia advinda de fontes renováveis, como a eólica, a solar e a hidroelétrica.

Concorrentes Atuais: Os concorrentes são aqueles que ofertam produtos iguais ou semelhantes aos oferecidos pela empresa em questão. Desta forma, os concorrentes atuais são empresas que, semelhantemente à PowerCo, fornecem energia não renovável na Europa. Dentre essas empresas pode-se destacar a Total Energies, a SSE Airtricity e a BP.

Novos Concorrentes: Os novos concorrentes são organizações de recente entrada no mercado, as quais possuem produtos e público alvo semelhantes a de sua empresa. Logo, como novo concorrente vale citar a aliança feita entre os grandes nomes do ramo energético na Europa, a qual visa impulsionar uma revolução energética verde que atinja todo o continente. Tal fator culminaria na diminuição de clientes da PowerCo, a qual utiliza uma matriz energética não renovável e, consequentemente, não é parte da aliança supracitada.

Poder de Negociação do Fornecedor: Em relação aos fornecedores de recursos para a PowerCo, cabe destacar o setor de gás natural e o setor de petróleo. A guerra entre a Ucrânia e a Rússia tem afetado diretamente o mercado energético na Europa, dado que a Rússia se consolidava como um dos principais fornecedores de gás natural e petróleo deste continente até o final de janeiro de 2022, fornecendo tais recursos para diversas empresas do ramo, entre elas a PowerCo. Com a redução do abastecimento russo, que representava 31% das importações europeias de petróleo bruto até o final de janeiro de 2022 (segundo o Serviço de Estatística da União Europeia), a empresa pode passar por dificuldades na obtenção de insumos para a produção de energia.

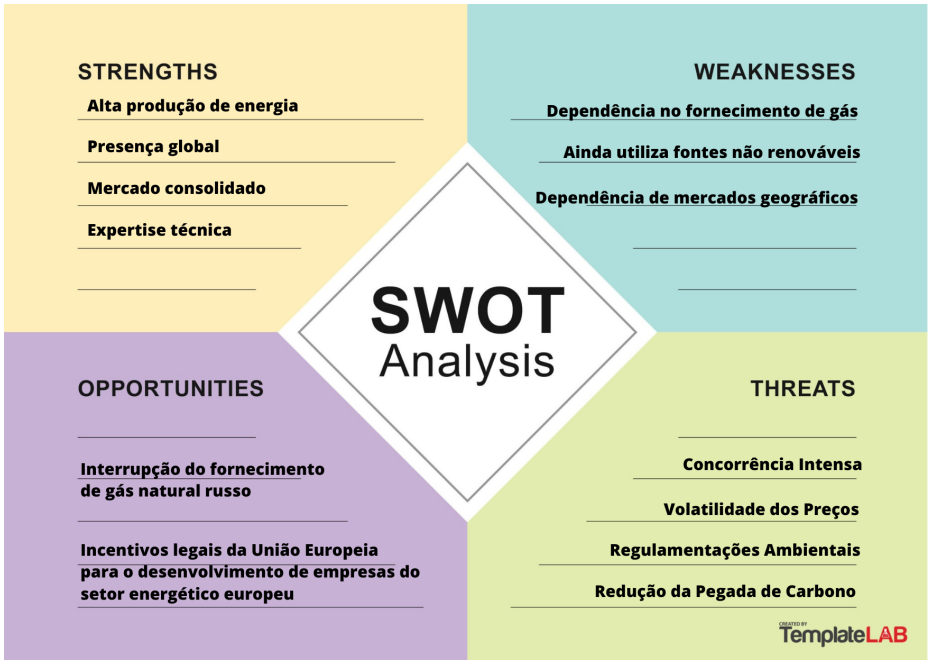
Poder de Negociação do Cliente: Por conta de temas como sustentabilidade e redução de impacto ambiental, o mercado atualmente depara-se com uma grande demanda vinda dos clientes: o fornecimento de energia limpa advinda de fontes renováveis. No entanto, em decorrência dos cortes no suprimento de gás natural à Europa por parte da Rússia, a PowerCo tem cogitado a utilização de usinas termoelétricas à base de carvão. Tal fator, é capaz de impactar o churn de clientes da PowerCo e consequentemente reduzir a lucratividade da empresa.

4.1.2. Análise SWOT

De forma complementar à análise de fatores "macro" da 5 Forças de Porter, a análise SWOT contempla fatores "micro" da empresa examinada, o que inclui suas *forças* (Strengths), *fraquezas* (Weaknesses), *oportunidades* (Opportunities) e *ameaças* (Threats). Nesse sentido, é possível definir as forças e as oportunidades como fatores de impacto *positivo* na instituição analisada, sendo a diferença entre elas dada pelo fato de representarem, respectivamente, fatores manipuláveis pela empresa e fatores externos ao controle desta. A mesma definição cabe às fraquezas e às ameaças, que se enquadram como fatores de impacto negativo (respectivamente) dentro e fora de controle da empresa.

Com isso em vista, a partir da figura a seguir, foi elaborada uma análise SWOT sobre a PowerCo:

Quadro 2 - Análise SWOT



Fonte: Autoria própria

- **Forças:**
  - Na análise SWOT, as forças são caracterizadas como fatores internos à empresa e que apresentam-se como diferenciais quando compara-se esta à concorrência. Assim, como características benéficas próprias da PowerCo e dos produtos que ela oferece, vale destacar sua estabilidade no mercado consolidado, uma vez que o mercado de energia europeu é bem estruturado e a PowerCo, até o momento, se mostrou bem integrada à estrutura do mercado por meio do fornecimento de gás natural e energia elétrica de fontes diversificadas, o que dificulta seu enfraquecimento na indústria energética. Além disso, tanto por conta de sua presença e manutenção no mercado consolidado, é inegável que a empresa seja caracterizada pela sua expertise técnica, essencial para a produção e fornecimento de gás e energia derivada de várias fontes, o que provoca uma maior confiança na companhia por parte dos consumidores ou clientes em potencial.
- **Oportunidades:**
  - As oportunidades, são definidas como eventos e/ou elementos benéficos para a empresa que não são essencialmente causados ou controlados por esta. Logo, as oportunidades da PowerCo, no momento atual, se concentram em dois principais tópicos. O primeiro é referente à Guerra Russo-Ucraniana, que deu origem a um grande corte no fornecimento de gás da Rússia aos países europeus — segundo artigo do portal jornalístico CNN, o fornecimento de gás da Rússia para a Europa caiu de 45% para 17,4% entre 2021 e 2023 —, fazendo com que vários consumidores tivessem que recorrer a novos fornecedores de energia (como a PowerCo). O segundo refere-se aos incentivos legais por entidades governamentais da União Europeia para o

crescimento de companhias de energia de países do bloco, os quais se deram como resposta ao corte de fornecimento supracitado e dão margem para a expansão da PowerCo sob o pretexto de alcançar a autonomia energética europeia.

- **Fraquezas:**
  - As fraquezas são fatores internos à empresa que podem configurar-se, de algum modo, como malefícios a esta. Como fraqueza, é imprescindível levar em conta a reação da companhia ao incentivo da União Europeia ao uso de fontes de energia renováveis ou limpas, uma vez que isso afeta diretamente todas as empresas do ramo energético no continente, mas especialmente a PowerCo, tendo em vista que sua matriz energética ainda é composta, em grande parte, por fontes de energia fósseis. Isto, somado ao baixo alcance da empresa a locais geologicamente propícios à obtenção de energia renovável (como planícies para energia eólica ou solar, por exemplo), põe em evidência, novamente, um mal preparo para a nova onda de energia limpa na Europa.
- **Ameaças:**
  - As ameaças são elementos externos à companhia que podem apresentar riscos aos seus interesses. Entre as ameaças enfrentadas pela PowerCo, deve-se considerar a alta concorrência em seu setor de atuação. Isso decorre do caráter consolidado do mercado de energia europeu, o qual significa que a maior parte das companhias concorrentes já possuem sua reputação estruturada e dificilmente serão abaladas drasticamente de forma a favorecer a PowerCo. Ademais, há a questão socioambiental, a qual, sob a ótica da crescente conscientização sobre o tópico de energia limpa, dificulta a manutenção do fornecimento de energia fóssil com a mesma intensidade de tempos passados, o que afeta diretamente o planejamento de suas usinas subsidiárias baseadas na extração de gás ou na utilização de carvão para geração termoelétrica.

4.1.3. Planejamento Geral da Solução

O planejamento geral da solução consiste, inicialmente, no processo de definir os dados disponíveis em relação ao problema apresentado e em como a solução será elaborada a partir destes. Portanto, a seguir, define-se também qual será a solução proposta, a tarefa executada por esta, como será utilizada, os benefícios oriundos de seu uso e os critérios adotados para definir seu sucesso.

a) **Dados disponíveis**

Para o desenvolvimento da solução, foi disponibilizada uma grande variedade e quantidade de dados oriundos dos próprios bancos de dados da PowerCo, os quais contemplam desde informações mais diretamente relacionadas aos clientes (como a quantidade de anos em vínculo com a empresa) até dados relacionados ao próprio consumo de energia (como a quantidade mensal de energia consumida em kWh). Contudo, os dados também apresentam um número considerável de inconsistências, o que torna o pré-processamento destes imprescindível para o desenvolvimento pleno da solução.

b) **Solução proposta**

Dito isso, a solução se enquadra, inicialmente, como um modelo preditivo capaz de informar se determinados clientes tendem ou não a realizar churn (ou, em outras palavras, desligar seu vínculo com a empresa). Em adição, também é esperado que a equipe de desenvolvimento torne o modelo capaz de indicar as melhores ações a serem tomadas para reduzir a probabilidade de churn de determinados clientes.

c) **Tipo de tarefa**

Portanto, percebe-se que a tarefa a ser realizada pelo modelo se caracteriza como classificação, uma vez que classificará os clientes como “propensos ao churn” ou “não propensos ao churn”.

d) **Modo de utilização da solução proposta**

A solução será utilizada, principalmente, pelo gerente de contas da PowerCo. Sua utilização será feita de modo a ajudar tal colaborador a identificar quais de seus clientes tendem a desligar seu vínculo com a empresa e, a partir desta informação, traçar medidas estratégicas para evitar o desligamento.

e) **Benefícios trazidos pela solução proposta.**

Diante do fato de que a predição em questão será feita através da tecnologia de máquina e de dados históricos em vez do pensamento e do empirismo humano, nota-se como benefício da solução a possibilidade de se realizar uma grande quantidade de previsões em um curto período de tempo e uma maior exatidão. Tais fatores retornarão para a empresa praticidade, economia de tempo, maximização de receita e maior segurança para a tomada de decisões, tendo em vista que impedir o churn de um cliente é menos custoso que revertê-lo.

f) **Critério de sucesso e métrica utilizada para avaliação**

Por fim, para avaliar o sucesso do modelo preditivo, a AUC-ROC das predições será usada como critério. Isto é, está estabelecido como métrica que, quanto maior for a AUC-ROC e quanto mais certeiras forem as precisões (evitando falsos negativos e falsos positivos), melhor sucedida será a solução no que tange a resolução do problema que este projeto se propõe a combater.

4.1.4. Value Proposition Canvas

O *Canvas Proposta de Valor* (ou Value Proposition Canvas), é uma ferramenta criada por Alex Osterwalder e usada para definir, com auxílio visual, o valor que determinado produto tem a oferecer para seu público alvo. Para isso, leva-se em conta diversos fatores relacionados aos clientes e diversos fatores relacionados ao produto. Este framework pode ser dividido em duas partes, uma que referencia-se ao cliente e outra que se referencia ao produto, além disso ambas possuem subdivisões, as quais serão apresentadas abaixo.

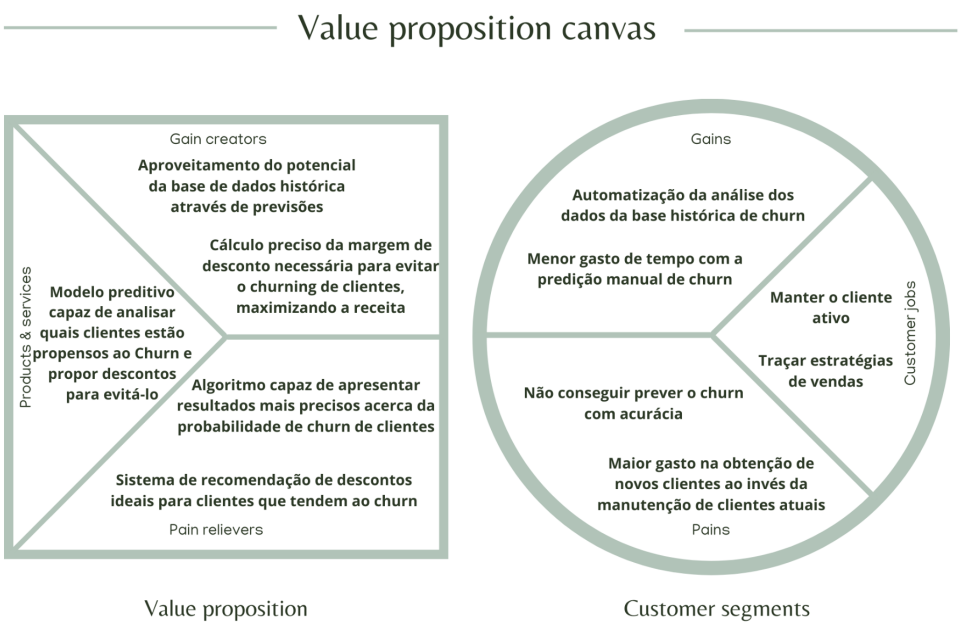
No que diz respeito à parte dos clientes, nela encontramos *customer jobs*, que são as tarefas referentes ao trabalho que os clientes precisam realizar no dia-a-dia, *pains*, que são

as “dores” ou dificuldades enfrentadas pelos clientes ao realizar tais atividades e *gains*, que podem ser descritos como os “ganhos” ou bonificações (de qualquer tipo) esperados pelos clientes ao cumprirem com sucesso suas tarefas.

Já entre os fatores referentes ao produto em si, há *products & services*, que trata da própria definição do produto e/ou serviço em questão, *pain relievers*, que consistem nos atributos do produto que podem aliviar as dificuldades enfrentadas pelo público alvo na realização de suas tarefas e *gain creators*, que referem-se aos atributos do produto capazes de atender aos ganhos almejados pelos clientes.

Diante disso, a equipe ForeSee elaborou um Value Proposition Canvas sobre o modelo preditivo em desenvolvimento — Forecast:

Quadro 3 - Value Proposition Canvas



Fonte: Autoria própria

Customer segments

- *Gains / Ganhos*
  - Entre os ganhos para a PowerCo, podemos citar a automatização da análise dos dados da base histórica de churn, processo que atualmente ainda é manual e demorado. Com isto, a PowerCo ganhará agilidade em seus processos e poderá redirecionar o tempo de seus colaboradores para tarefas essencialmente realizadas por humanos.

- *Customer jobs / Tarefas do cliente*
  - Entre as tarefas que precisam ser realizadas pelos funcionários da PowerCo, cabe destacar a manutenção dos clientes ativos da empresa, o que se relaciona diretamente com a criação de estratégias de vendas, as quais envolvem a definição de ações ideais para minimizar a evasão de clientes.
- *Pains / Dores*
  - As maiores dores enfrentadas pela PowerCo concentram-se na baixa acurácia na previsão de churn dos clientes. Isso se dá em decorrência de, atualmente, não existir nenhuma tecnologia diferente da própria análise manual humana para guiar a realização desse processo na empresa, o que resulta numa taxa maior de churn. Dessa forma, a margem de lucro da PowerCo reduz, uma vez que a obtenção de novos clientes para compensar a perda de clientes antigos é mais custosa que a manutenção de clientes já estabelecidos na companhia.

Value Proposition

- *Products & services / Produtos e serviços*
  - Para trazer a proposta de valor, nosso produto é um Modelo Preditivo que utiliza Machine Learning para, através da análise do banco de dados da PowerCo, retornar quais clientes estão mais propensos a um turnover dos serviços da empresa. Além disso, através de cálculos de preços e outras variáveis, nosso produto retornará uma porcentagem ideal de desconto e/ou outras medidas para a empresa reduzir o número de clientes com alta probabilidade de dar churn.
- *Gain creators / Criadores de ganho*
  - Para um melhor aproveitamento dos ganhos já existentes dentro da PowerCo, nosso produto utilizará da grande base de dados que a empresa já apresenta e trará uma maior utilidade para ela. O algoritmo identifica, através de análises no banco de dados, os clientes em que a empresa deverá tomar alguma medida da forma estatisticamente mais viável para a empresa em questões de lucratividade.
- *Pain relievers / Aliviadores de dor*
  - Para os problemas que a PowerCo sofre, como a baixa acurácia nas previsões manuais de churn e gastos para obtenção de novos clientes, nosso produto é capaz de trazer resultados mais precisos por conta do aprendizado de máquina. Isso, em conjunto com o sistema de recomendação que propõe valores ótimos de desconto para clientes que tendem ao churn, terá como consequência uma menor taxa de churn na PowerCo, fator que acarreta em uma menor necessidade de esforços da empresa na tentativa de obtenção de novos clientes.

4.1.5. Matriz de Riscos

A *matriz de risco*, segundo o artigo homônimo da Universidade Federal da Integração Latino Americana, é uma ferramenta visual para analisar os riscos envolvidos dentro de um projeto. Ela aborda tanto os riscos negativos quanto os riscos positivos existentes no

desenvolvimento do projeto em questão, exibindo-os por meio de uma tabela cujos eixos representam a probabilidade de ocorrência do risco e o nível de impacto deste sobre o projeto. Tendo em vista o mapeamento e a contingência de riscos que a matriz de risco proporciona, a ForeSee elaborou uma a respeito do Forecast, a qual pode ser visualizada logo abaixo:

Quadro 4 - Matriz de Riscos

	Riscos Negativos			Riscos Positivos			
90%	Ocorrer vazamento de dados privados	Não terminar o projeto a tempo			Atender as expectativas do cliente em relação ao projeto		90%
60%	Dados não suficientes para uma análise aprofundada	Não conseguir realizar tratamento de dados eficiente	Membros do grupo possuírem baixa frequência em dailies e devs	A empresa gostar do projeto a ponto de dar continuidade	Finalizar o projeto com antecedência e adicionar mais funcionalidades		60%
30%			O grupo, em geral, não entender como utilizar a biblioteca Panda's	Consistência geral dos dados	O treinamento do Modelo Preditivo dar resultados satisfatórios com poucas tentativas	Projeto ter o escopo reduzido	30%
P/I	Baixo	Médio	Alto	Alto	Médio	Baixo	

Fonte: Autoria própria

Riscos negativos

- **Ocorrer vazamento de dados privados**
  - Tendo em vista que o projeto será feito com o uso de plataformas online (como o GitHub e o Google Colab) para confecção e armazenamento de código, dados e documentação, há uma possibilidade significativa de vazamento de dados particulares da PowerCo. Contudo, a equipe ForeSee está preparada para evitar que tal vazamento ocorra, haja vista que já possuem as informações a respeito de quais elementos do projeto poderão ou não ser públicos, bem como orientações de como lidar com estes.
- **Não terminar o projeto a tempo**
  - Existe uma chance considerável de que o MVP do projeto não seja concluído a tempo. Isso se dá por conta, principalmente, da possibilidade de uma eventual má organização ou de um eventual mal aproveitamento de tempo por parte da equipe. Logo, a ForeSee está empenhada em manter o progresso do projeto bem organizado e visível através de Kanban e dailies.
- **Dados não suficientes para uma análise aprofundada**
  - As bases de dados que constam entre os arquivos fornecidos para o desenvolvimento do projeto, ao menos à primeira vista, aparentam serem insuficientes, haja em consideração a grande quantidade de campos de dados vazios e/ou preenchidos com “NaN”. Contudo, espera-se que, com o pré-processamento de dados, grande parte dos dados incompletos possam

ser manipulados ou até inutilizados de forma que o modelo preditivo possa ser desenvolvido eficientemente.

- **Não conseguir realizar tratamento de dados eficiente**
  - A possibilidade de que os dados não sejam tratados adequadamente tem sua parcela de significância entre os riscos em decorrência da importância da boa execução dessa etapa do projeto para que as todas as demais sejam igualmente bem executadas. Contudo, há chances de que o tratamento de dados tenha falhas relevantes, uma vez que é a primeira experiência de todos os membros da equipe trabalhando diretamente com inteligência artificial e machine learning.
- **Membros do grupo possuírem baixa frequência em dailies e devs**
  - A presença de membros da equipe em dailies e devs tem um impacto imenso no desenvolvimento do projeto, uma vez que o nível de participação das pessoas nesses momentos impacta não só na organização do grupo, mas também diretamente na própria confecção do modelo preditivo. Dessa forma, caso essa participação não seja constante, todos os demais riscos negativos terão maior probabilidade de acontecimento, o que só pode ser contido através da autodisciplina e responsabilidade de cada integrante da ForeSee.
- **O grupo, em geral, não entender como utilizar a biblioteca Pandas**
  - Existe uma pequena chance de que os membros da equipe não entendam o funcionamento da biblioteca Pandas adequadamente, o que dificultaria sua utilização e, portanto, o desenvolvimento do projeto em si. Contudo, com os encontros de instrução, autoestudos e apoio dos professores de programação, por mais que exista pouca experiência do time ForeSee com a utilização do Pandas, o aprendizado sobre a biblioteca se apresenta como totalmente viável e bem direcionado. Portanto, esse risco é contornável e possui baixíssima relevância no momento.

Riscos positivos

- **Atender as expectativas do cliente em relação ao projeto**
  - Apesar de todos os riscos negativos, levando em conta o histórico de entregas finais de projetos passados dos integrantes da ForeSee, há uma grande chance de que o projeto seja entregue de forma coerente com as expectativas do cliente. Isso se confirma por meio da capacidade — que os membros possuem — de rápido planejamento de medidas para contornar as dificuldades encontradas ao longo do projeto.
- **A empresa gostar do projeto a ponto de dar continuidade**
  - Levando em conta o risco benéfico anterior, há a possibilidade de que a entrega final do projeto agrade o cliente a ponto deste decidir dar continuidade com o desenvolvimento e aprimoramento do projeto.
- **Finalizar o projeto com antecedência e adicionar mais funcionalidades**
  - Ainda considerando os riscos positivos supracitados, é possível que as funcionalidades principais que constam no MVP sejam desenvolvidas e implementadas e, por consequência, haja uma abertura de tempo para inserir

funcionalidades extras no modelo preditivo. Contudo, é importante que a equipe foque, pelo menos no momento, em seguir apenas os requisitos mínimos exigidos dentro do escopo do projeto.

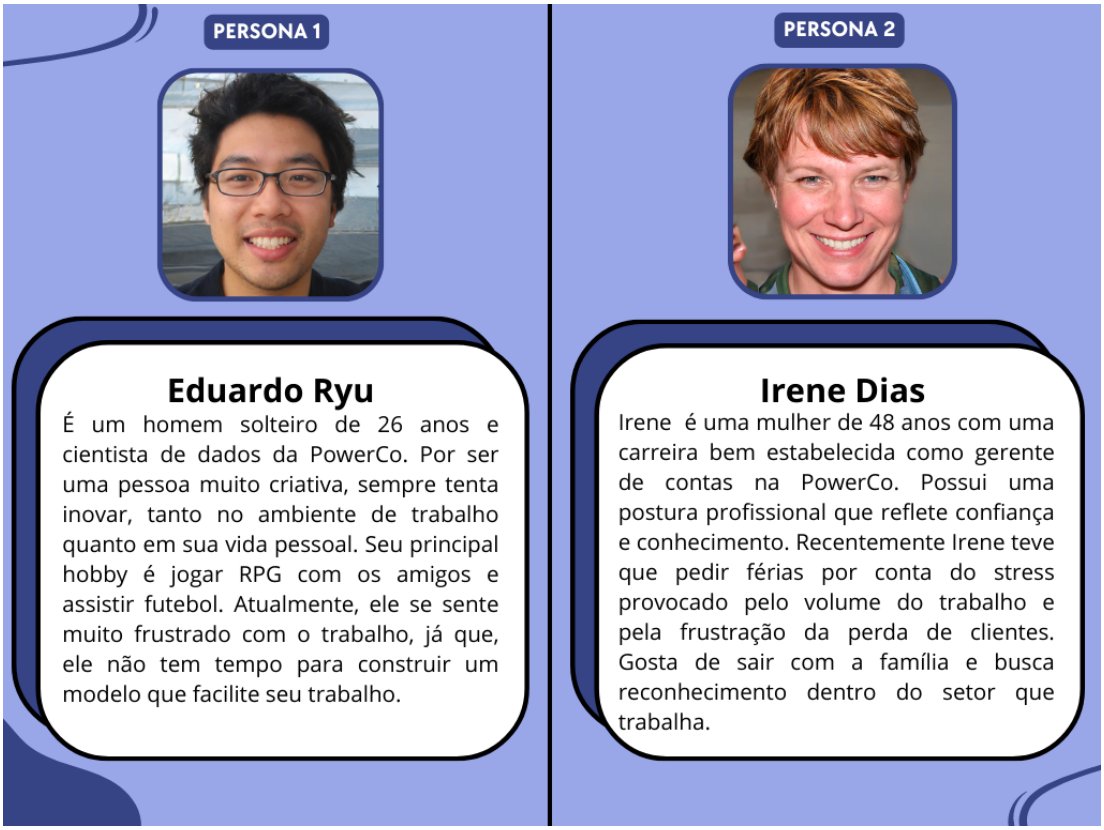
- **Consistência geral dos dados**
  - Como as bases de dados ainda não foram completamente exploradas pelo time de desenvolvimento, não há uma certeza definitiva sobre a consistência dos dados que constam nelas, mas existem esperanças de que estes apresentem uma coerência e coesão mínima, o que facilitaria o trabalho de pré-processamento de dados. Contudo, é nítido que muitos dos dados possuem inconsistências ou faltas, o que leva o risco positivo em questão para um pouco mais longe da realidade.
- **O treinamento do Modelo Preditivo dar resultados satisfatórios com poucas tentativas**
  - Existe a possibilidade de que o treinamento do modelo preditivo seja feito de forma rápida, com predições satisfatórias em poucas tentativas. Contudo, isso é algo pouco provável de acontecer, uma vez que o treinamento é um processo intensamente iterativo e, ainda, que nossa equipe não possui grande experiência com esse tipo de processo.
- **Projeto ter o escopo reduzido**
  - Existe, ainda, uma pequena chance de o escopo do projeto ser reduzido durante o decorrer de seu tempo de desenvolvimento, o que poderia facilitar e agilizar o trabalho do time ForeSee. Entretanto, o escopo do projeto já foi bem definido e estruturado já na primeira semana de trabalho, o que torna a probabilidade da concretização desse rico positivo quase nula.

4.1.6. Personas

As *personas*, de acordo com a Interaction Design Foundation, podem ser descritas como perfis específicos de pessoas contempladas pelo público alvo de determinado produto em desenvolvimento. Para além de descrições genéricas de um perfil amplo de clientes, as personas abordam indivíduos em suas particularidades, atribuindo-lhes nomes, características psicológicas, detalhes sobre suas atividades desenvolvidas e seu comportamento. Sua utilidade advém do vínculo mais íntimo construído entre o desenvolvedor do produto e o usuário em potencial, permitindo que este primeiro consiga compreender — de forma mais razoável — a visão de mundo do segundo e o impacto dela sobre a utilização de seu produto.

Com isso em mente, a seguir, foram elaboradas duas personas referentes ao Forecast:

Quadro 5 - Personas



Fonte: Autoria própria

Além disso, o Eduardo Ryu é um cientista de dados que trabalha com Python. Por ser muito criativo, ele acaba sendo um tanto desorganizado. Já a Irene Dias, por ser um pouco mais madura, é totalmente organizada e domina a ferramenta do excel. Seu maior ponto forte é o atendimento ao cliente.

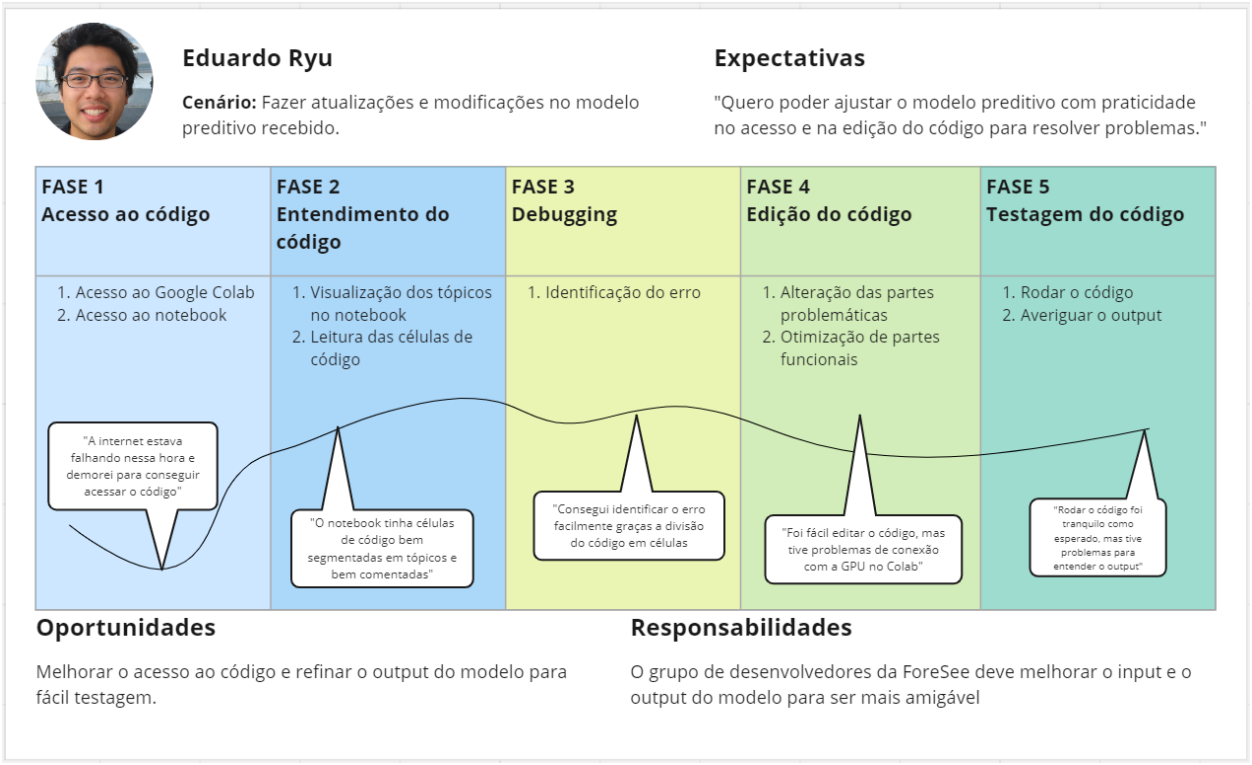
À vista das personas elaboradas, entende-se que o ForeCast deve ser desenvolvido de forma a considerar um escopo duplo de público alvo: pessoas com conhecimento técnico da área de ciência de dados e pessoas da área de negócios que usufruem das tecnologias desta outra área. Por isso, é importante pensar, desde já, em formas de tornar o uso da ferramenta amigável para ambos os usuários, evitando o emprego de termos técnicos ou interfaces não intuitivas para pessoas fora da área de ciência de dados, mas mantendo a possibilidade de ajustes e aprimoramentos práticos do modelo por integrantes da área de tecnologia da informação.

4.1.7. Jornadas do Usuário

De acordo com a MJV Tecnologia & Inovações, as *jornadas de usuário* são, de forma sucinta, o processo pelo qual um usuário passa desde o início até o fim da utilização de um produto. Elas se diferenciam das “user stories” (ou histórias de usuário) por apresentarem, além de descrições do processo em si, as emoções e pensamentos em potencial que os usuários tendem a desenvolver ao longo de cada uma das etapas do manuseio do produto.

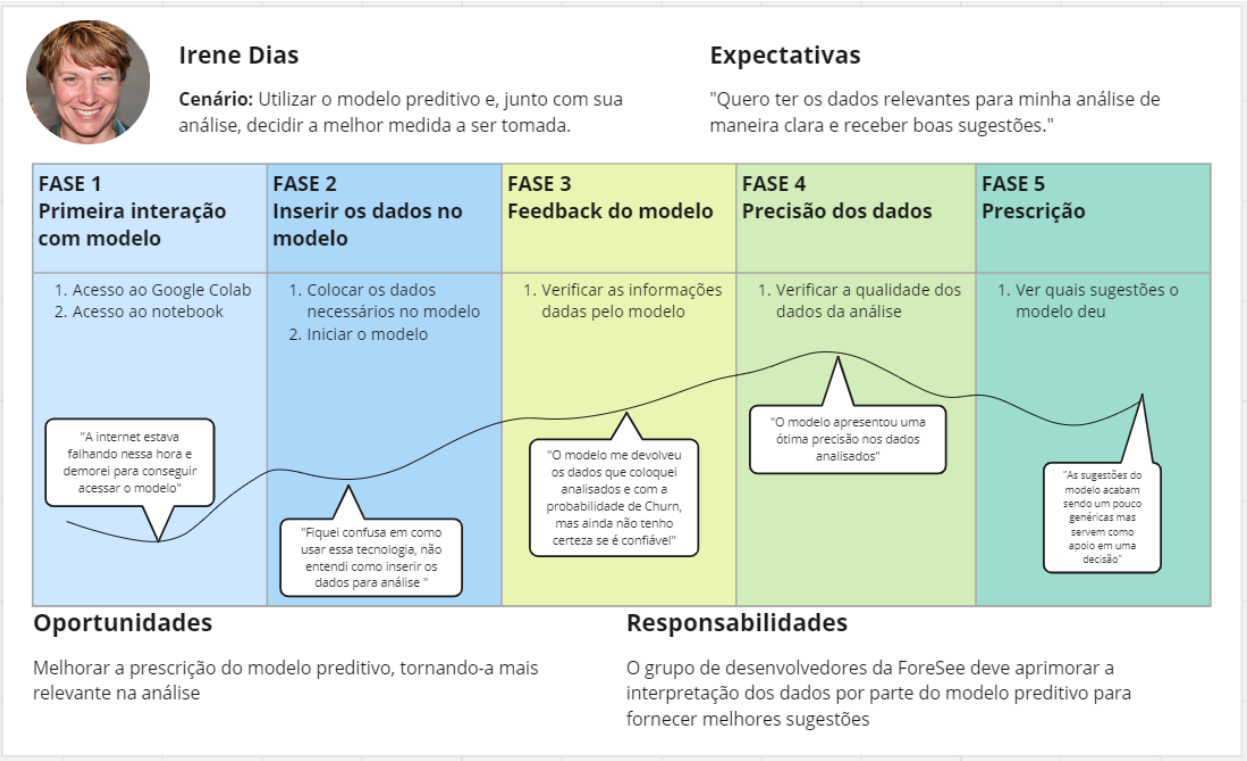
Visando, novamente, uma conexão maior com a visão de mundo dos potenciais usuários do Forecast, a ForeSee elaborou duas jornadas de usuário com base nas personas apresentadas na seção anterior.

Quadro 6 - Jornada de Usuário 1: Eduardo Ryu



Fonte: Autoria própria

Quadro 7 - Jornada de Usuário 2: Irene Dias



Fonte: Autoria própria

Através da análise tem-se como objetivo canalizar esforços para melhorar a solução visando as dores que o cliente tem ao usar o produto, como observado nas abas de oportunidades e responsabilidades. Dentre estas, as responsabilidades têm prioridade a serem realizadas por serem partes necessárias do produto, já as oportunidades são aditivos para que o usuário supere suas expectativas.

4.1.8 Política de Privacidade

A *política de privacidade* compõe uma parte fundamental do desenvolvimento do Forecast, haja em vista a necessidade de atender à regulamentação imposta pela LGPD (Lei Geral de Proteção de Dados). Por meio da transparência sobre quais dados estão sendo utilizados, como eles são manipulados e com quem eles são compartilhados no contexto de desenvolvimento do Forecast, é possível estabelecer uma relação ética e de confiança entre desenvolvedor, produto e usuário. Por isso, a seguir, há uma versão sucinta da política de privacidade adotada pela PowerCo, a qual será usada como guia no desenvolvimento do presente projeto.

Informações sobre o tratamento de dados 1. *Quais dados pessoais são coletados (inclusive os dados não informados pelo usuário, como IP, localização, etc)?*

- Os dados coletados são os de nome, email, dados de conta bancária, telefone e endereço.

2. *Onde os dados são coletados (fonte)?*

- Os dados foram coletados durante o processo de contratação do serviço.

3. *Para quais finalidades os dados são utilizados?*
    - Eles são utilizados para coleta, transformação dos dados e análise estatística.
4. *Onde os dados ficam armazenados?*
    - Eles estão localizados no banco de dados da PowerCo e em arquivos CSV para análise.
5. *Qual o período de armazenamento dos dados (retenção)?*
    - A retenção dos dados é permanente.
6. *Uso de cookies e/ou tecnologias semelhantes? Sobre este processamento*
    - Em conexão com o uso de cookies tecnicamente necessários, processamos seus dados pessoais para fornecer a você certas funções deste site.

Descrição e finalidade do processamento

Quando você acessa este site, seus dados pessoais podem ser processados em conexão com o uso dos chamados "cookies". Cookies são arquivos de texto armazenados em seu dispositivo terminal e são necessários para determinadas funções deste site. Você pode definir as configurações do seu navegador de acordo com suas preferências e, por exemplo, recusar-se a aceitar cookies de terceiros ou todos os cookies. Observe que talvez você não consiga usar todas as funções deste site. A este respeito, a seguinte informação refere-se a cookies que não têm ligação a um serviço disponível através do nosso site, mas servem exclusivamente para fornecer o próprio site ("cookies tecnicamente necessários").

Categorias de dados pessoais

Gerenciador de tags do Google: dados agregados sobre o acionamento de tags.

Content Management Platform Usercentrics: dados de inclusão e exclusão, URL de referência, agente do usuário, configurações do usuário, ID de consentimento, horário e tipo de consentimento, versão do modelo, idioma do banner.

Base legal do processamento

O processamento de seus dados pessoais em conexão com o uso de cookies tecnicamente necessários neste site é baseado em nosso interesse legítimo (Art. 6 (1) lit. f DSGVO). Equilibramos nosso interesse em fornecer as funções deste site que dependem de cookies com seu interesse na confidencialidade de seus dados pessoais, pelo que prevalece nosso interesse. Sem o processamento de dados pessoais, é tecnicamente impossível fornecer as funções. Ao mesmo tempo, você tem a opção descrita previamente para impedir o processamento de seus dados pessoais em conexão com cookies.

Destinatários / categorias de destinatários: Não divulgamos seus dados pessoais aos destinatários.

Período de armazenamento: Conservamos os seus dados pessoais durante o tempo necessário à utilização do respetivo cookie.

Obrigação de fornecer: Você não é obrigado a divulgar dados pessoais. Sem seus dados pessoais, nem todas as funções deste site podem estar disponíveis para você

7. *Com quem esses dados são compartilhados (parceiros, fornecedores, subcontratados)?*
   - Não serão compartilhados.
8. *Informações sobre medidas de segurança adotadas pela empresa.*
   - Não existe fonte sobre essa informação.
9. *Orientações sobre como a empresa/organização atende aos direitos dos usuários.*
   - Não existe fonte sobre essa informação.
10. *Informações sobre como o titular de dados pode solicitar e exercer os seus direitos.*

    O titular de dados pode entrar em contato com o Encarregado de Proteção de Dados por email e solicitar os devidos relatórios. Os direitos do titular são: - Direito de Acesso (Artigo 15 do GDPR) - Direito de Retificação (Artigo 16 do GDPR) - Direito ao Apagamento (ou "Direito de Ser Esquecido") (Artigo 17 do GDPR) - Direito à Portabilidade dos Dados (Artigo 20 do GDPR) - Direito de Oposição (Artigo 21 do GDPR) - Direito de Restrição de Processamento (Artigo 18 do GDPR) -Direito de Retirar o Consentimento (Artigo 7 do GDPR)

Responsabilidade pelos Dados: A E.ON SE é responsável pelo tratamento dos dados pessoais.

Acesso e Correção de Dados: Os indivíduos têm o direito de solicitar informações sobre os dados armazenados sobre eles e de corrigir dados incorretos.

Restrição, Portabilidade e Exclusão de Dados: Os indivíduos podem solicitar a restrição do processamento, a portabilidade dos dados fornecidos e a exclusão dos dados, desde que não sejam mais necessários.

Oposição ao Processamento: Os indivíduos têm o direito de se opor ao uso de seus dados com base em interesses públicos ou legítimos.

Retenção de Dados: Alguns dados podem ser retidos para garantir o funcionamento do site, e não é possível se opor ao processamento dessas informações.

Revogação de Consentimento: Se os dados forem processados com base no consentimento do indivíduo, esse consentimento pode ser revogado a qualquer momento.

Contato: Os indivíduos podem entrar em contato com a E.ON SE para exercer seus direitos ou fazer reclamações.

Autoridade Reguladora: Os indivíduos também têm o direito de encaminhar uma reclamação a uma autoridade reguladora, como o Comissário Estadual de Proteção de Dados e Liberdade de Informação North Rhine-Westphalia, na Alemanha.

11. *Informações de contato do Encarregado de Proteção de Dados (DPO - Data Protection Officer) da organização.*

    Data Protection Officer

    Brüsseler Platz 1

    D-45131 Essen

    datenschutz@eon.com

    https://www.eon.com/en/privacy.html

4.2. Compreensão dos Dados

4.2.1. Exploração de dados

Categorização dos Dados

Acerca da identificação dos dados das colunas, podemos classificá-los de duas maneiras distintas: numéricos ou categóricos. São categorizados como dados numéricos aqueles que representam quantidades, estando sempre na forma de números. Como categóricos, são especificados aqueles que representam grupos ou - assim como sua nomenclatura diz - categorias.

Dentro desses conceitos, há ainda subdivisões. Os dados numéricos ramificam-se em **Numéricos Discretos**, os quais assumem apenas valores inteiros, e **Numéricos Contínuos**, que adotam valores não inteiros, também chamados de 'quebrados'. Já os Categóricos, são subclassificados em **Categóricos Nominais**, que representam grupos ou características, e **Categóricos Ordinais**, os quais identificam categorias ordenadas por algum parâmetro.

A partir de tais conceitos são classificadas as colunas das três bases de dados oferecidas pelo cliente. Segue-se esta análise, em individual, para cada uma das bases de dados.

Análise "base_clientes.csv"

Dados numéricos discretos

- date_activ
- date_end
- date_first_activ
- date_modif_prod
- date_renewal
- nb_prod_act
- num_years_antig

Dados numéricos contínuos

- cons_12m
- cons_gas_12m
- cons_last_month
- forecast_base_bill_ele

- forecast_base_bill_year
- forecast_bill_12m
- forecast_cons_year
- forecast_discount_energy
- forecast_meter_rent_12m
- forecast_price_energy_p1
- forecast_price_energy_p2
- forecast_price_pow_p1
- imp_cons
- margin_gross_pow_ele
- margin_net_pow_ele
- net_margin
- pow_max

  Dados categóricos nominais
- id
- activity_new
- campaign_disc_ele
- has_gas
- origin_up
- channel_sales

  Dados categóricos ordinais

  -> Não há colunas com dados categóricos ordinais nessa base de dados

Análise “base_precos.csv”

Dados numéricos discretos

- price_date

  Dados numéricos contínuos
- price_p1_var
- price_p2_var
- price_p3_var
- price_p1_fix
- price_p2_fixr
- price_p3_fix

  Dados categóricos nominais
- id

  Dados categóricos ordinais

  -> Não há colunas com dados categóricos ordinais nessa base de dados

Análise “base_hist_churn.csv”

Na base “base_hist_churn”, há apenas duas colunas: ‘id’ e ‘churn’. Ambas as colunas possuem dados Categóricos Nominais.

Através da observação, é possível afirmar que os dados numéricos compõem a maior parte das colunas presentes nas nossas bases de dados. Desta forma, com a categorização dos dados, é possível entender os tipos destes, para, então, dar-lhes o tratamento mais apropriado de acordo com suas respectivas características.

Análise Estatística Descritiva

A análise estatística descritiva é amplamente utilizada para o entendimento de determinado conjunto de dados. Na ciência de dados, é imprescindível o uso desta para que possamos compreender os dados numéricos com os quais estamos trabalhando. A partir de tal análise, encontramos a média, a moda, a mediana, o desvio-padrão e outras medidas matemáticas que auxiliam no manejo dos dados.

Sendo assim, inicialmente, foi feito um estudo dos dados numéricos brutos contidos nas bases de dados, os quais sustentam o desenvolvimento do modelo preditivo. Contudo, com os resultados obtidos através desta análise, percebeu-se que os dados brutos não permitem uma estatística descritiva ideal, uma vez que, por exemplo, existem valores muito altos na linha “std”, que se refere ao desvio padrão - medida que representa um valor aproximado geral da diferença de cada valor da coluna em relação à média deles mesmos. Logo, é possível concluir que existem dados com valores muito destoantes da média, o que, de forma resumida, constitui uma grande quantidade de outliers.

A partir disto, é nítido que tais valores afetam a análise de forma significativa, tornando necessário, para que a estatística descritiva seja factível, tratar os dados. Então, a fim de obter uma melhor compreensão dos resultados da estatística descritiva e tirar conclusões mais confiáveis, os dados foram corretamente tratados - processo documentado na seção 4.2.2 -, o que possibilitou a análise a seguir:

- A primeira base a ser analisada foi a **“base_clientes.csv”**, na qual notou-se que os maiores valores de desvio-padrão estavam presentes nas seguintes colunas: **“cons_12m”**, **“cons_gas_12m”** e **“cons_last _month”** — chegando a apresentar variação de 6 dígitos.
  - -> cons_12m: 566344.729034
  - -> cons_gas_12m: 163188.505456
  - -> cons_last_month: 63651.557243
    - A coluna “cons_12m’ também apresentou média e mediana bastante elevadas, o que justifica o alto desvio padrão.
    - Já no caso das colunas “cons_gas_12m” e “cons_last _month”, a média e mediana são bem mais baixas, o que leva à conclusão de que a coluna possui uma alta discrepância entre os valores contidos nela.

- Na **“base_precos.csv”**, também foram analisadas todas as colunas. Com base no raciocínio anterior, as colunas **“price_p1_fix”**, **“price_p2_fix”** e **“price_p3_fix”** são as que apresentam as maiores médias e também as que apresentam os maiores desvios padrão.
  - -> price_p1_var: 0.024251
  - -> price_p2_var: 0.049823
  - -> price p3 var: 0.036202
  - -> price_p1_fix: 5.317825
  - -> price_p2_fix: 12.821517
  - -> price_p3_fix: 7.760806
    - Ademais, com exceção da coluna “price_p1_fix”, que possui uma mediana de 44.281745, todas as outras possuem uma mediana próxima de 0, demonstrando que a maioria dos preços pagos pelos clientes são de baixo valor.
    - **• (média)**
      - price_p1_var 0.141025
      - price_p2_ var 0.054322
      - price_p3_var 0.030692
      - price_p1_fix 43.335137
      - price_p2_fix 10.692625
      - price 03 11x 6.454092
    - **• (mediana)**
      - price_p1_var 0.147251
      - price_p2_var 0.085884
      - price_p3_var 0.000000
      - price_p1_fix 44.281745
      - price_p2_fix 0.000000
      - price_p3_fix 0.000000

Assim, foi possível compreender características importantes de nossos dados, como também extrair informações relevantes para a construção do modelo.

4.2.2. Pré-processamento dos dados

O pré-processamento dos dados, terceira etapa do CRISP-DM - metodologia utilizada durante o processo de construção do modelo -, é o momento onde, após analisarmos e entendermos nossa base, começamos a tratá-la a fim de prepará-la para ser utilizada na construção do modelo preditivo. Assim, são removidos valores nulos ou inconsistentes, colunas que não possuem relação com o que estamos abordando e certos dados são padronizados ou transformados, para que sejam melhor compreendidos pelo modelo que construiremos.

Exclusão de colunas e linhas inutilizáveis:

Com base nos dados ruidosos e/ou inconsistentes identificados na exploração de dados, o grupo operou um pré-processamento de dados, no qual ocorreu a deleção de algumas colunas das bases de dados. Em sua maioria, elas foram removidas por apresentar escassez de informações e, muitas vezes, campos de dados nulos, como nas colunas *"activity_new"* (categoria de atividade da companhia) e *"campaign_disc_ele"* (código de identificação da última companhia de eletricidade à qual o cliente se inscreveu), as quais, além de apresentarem esses problemas, diziam respeito a informações irrelevantes para a predição de churn feita pelo modelo e continham dados incoerentes, como preços negativos.

Lista de colunas excluídas

- activity_new
  - campaign_disc_ele
  - date_first_activ
  - forecast_cons

De forma complementar, a fim de reduzir a quantidade de informação inutilizável nas bases de dados, também foram excluídas linhas específicas que apresentavam valores nulos ou ruidosos (como preços negativos) em determinadas colunas. Segue uma lista de colunas cujos valores foram levados em conta nesse processo:

Lista de colunas consideradas

- date_end
  - date_modif_prod
  - date_renewal
  - forecast_discount_energy
  - forecast_price_energy_p1
  - forecast_price_energy_p2
  - forecast_price_pow_p1
  - margin_gross_pow_ele
  - margin_net_pow_ele
  - net_margin
  - origin_up
  - pow_max

Tratamento da base de dados de preços:

Além disso, a partir da exploração de dados, também notou-se uma inconsistência em relação aos dados contidos na base **"base_precos.csv"**. Nessa tabela, as colunas de *"price_pX_fix"* e *"price_pX_var"* (em que X representa o número do período da coluna em questão), indicadoras do preço pago em determinado período do dia por determinado

cliente em determinada data, apresentavam vários valores diferentes para mesmos clientes em diferentes datas, mas muito próximos entre si. Dessa forma, optou-se por fazer uma média dos valores que se repetiam e contê-la dentro de linhas únicas, o que levou a coluna *"price_date"* a ser dispensada, uma vez que o diferenciamento de datas se tornou não mais relevante.

PCA A Análise de Componentes Principais (PCA) é uma técnica estatística multivariada amplamente utilizada para reduzir a dimensionalidade de conjuntos de dados complexos. Ela foi desenvolvida por Karl Pearson, matemático inglês, e tem sido uma ferramenta valiosa em diversas áreas, como estatística, análise de dados, aprendizado de máquina e ciência da computação.

Fundamentação Teórica Segundo Pearson, o PCA tem como objetivo identificar a variância entre cada componente, vetores ortogonais uns aos outros, encontrando as direções ao longo dos quais os dados se espalham da forma mais ampla possível. Geralmente, as primeiras componentes principais contêm a maior parte da informação relevante nos dados originais.

O processo de PCA envolve os seguintes passos:

- **Padronização dos Dados:** Os dados originais são geralmente padronizados, o que significa que as variáveis são ajustadas para terem média zero e desvio padrão unitário. Isso é importante para garantir que todas as variáveis tenham a mesma escala.
- **Cálculo da Matriz de Covariância ou Correlação:** A matriz de covariância (ou correlação) é calculada a partir dos dados padronizados. Ela descreve como as variáveis estão relacionadas umas com as outras.
- **Decomposição da Matriz:** A matriz de covariância é então diagonalizada para obter os autovetores e autovalores. Os autovetores representam as direções dos componentes principais, enquanto os autovalores indicam a quantidade de variância explicada por cada componente.
- **Seleção das Componentes Principais:** As primeiras componentes principais são escolhidas com base nos autovalores, começando pela componente com o maior autovalor. Normalmente, decide-se um limite para a quantidade de variância explicada que se deseja reter.
- **Projeção dos Dados:** Os dados originais são projetados nas componentes principais selecionadas, resultando em um novo conjunto de dados com dimensões reduzidas.

Aplicações da PCA A PCA, com suas contribuições de Karl Pearson (Pearson, 1901), tem várias aplicações práticas, sendo a principal a redução de dimensionalidade. O PCA foi usado no projeto para reduzir a quantidade de variáveis no conjuntos de dados da PowerCo, visto que há uma alta dimensionalidade, identificando os componentes principais dentro das features e reduzindo esse número, facilitando a visualização e a análise.

Com isso, das 40 features aproximadas, sem contar com o processo de One Hot Encoding, a variável 'n' de número de componentes desejados foi estabelecido como 20. Portanto, ele

seleciona as 20 colunas que mais apresentam importância e expressa o dataset com apenas estas dimensões.

Outros pontos relevantes e conclusões:

As bases de dados também foram analisadas com foco na busca de outliers, os quais foram identificados, mas mantidos em razão de, no caso específico desse projeto, seu baixo impacto na acurácia do modelo preditivo e da baixa quantidade de linhas restantes no caso de deleção das linhas com dados numéricos muito dispersos em relação à média.

Outro ponto importante do pré-processamento de dados refere-se à quantidade de informação: antes do processamento de dados, a base de dados estava dividida em 3 **"base_precos.csv"**, **"base_clientes.csv"** e **"base_hist_churn"** — a primeira apresentava mais de 200 mil linhas, enquanto a segunda apresentava mais de 20 mil. Ao fim do pré-processamento, todas as tabelas foram unidas em uma única que, com a limpeza de dados já concluída, apresenta, atualmente, pouco menos de 15 mil linhas.

Ademais, com a exploração de dados antes do pré-processamento, ficou evidente que a quantidade de linhas de dados referentes a clientes que não cometeram churn era bem maior do que a de clientes que cometeram churn. Aproximadamente, 90.1% dos dados disponíveis diziam respeito apenas a clientes que não cometeram churn. Logo, a fim de obter dados mais diversos para manter uma maior acurácia do modelo preditivo, a equipe de desenvolvimento optou por utilizar a biblioteca **SMOTE** para balancear os dados por meio da interpolação de dados. Dessa forma, a base de dados apresenta uma distribuição de linhas de dados igualmente dividida (50%/50%) entre clientes que cometeram churn e clientes que não cometeram churn.

Com isso, a continuidade das próximas etapas de desenvolvimento do modelo preditivo torna-se mais prática para a equipe de desenvolvimento da ForeSee, reduzindo drasticamente problemas relacionados a dados não pré-processados.

4.2.3. Hipóteses

A partir do processo de entendimento dos dados, levantam-se hipóteses, as quais objetivam compreender as relações existentes entre determinadas colunas da base de dados. Após isto, tais hipóteses são averiguadas, através de gráficos ou outras análises. Se estas forem aceitas, servirão de base ou auxílio para o desenvolvimento do modelo. Se tal não ocorrer, são, então, levantadas novas hipóteses, as quais passarão por este processo até que sejam encontradas hipóteses que não podem ser negadas.

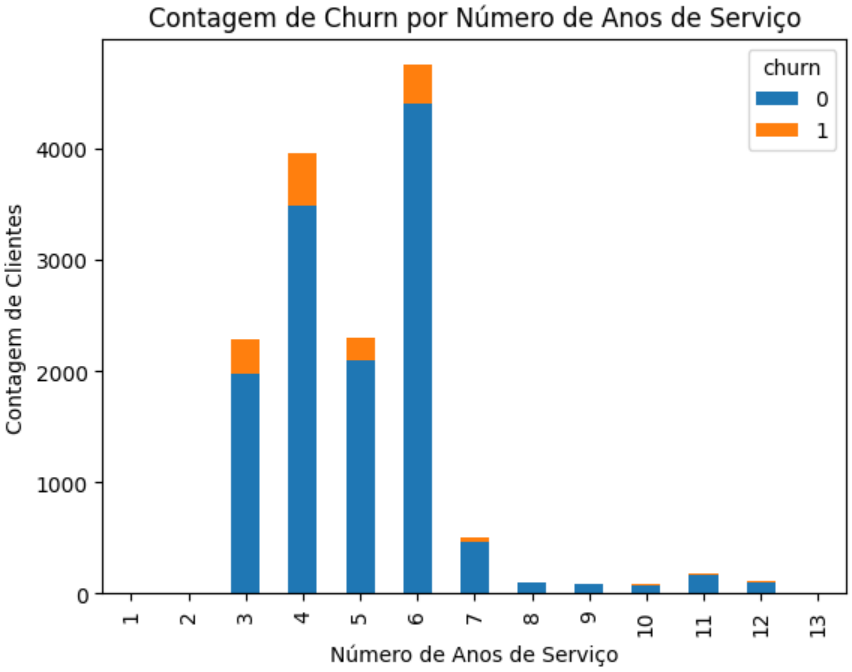
Sendo assim, para que a base de dados oferecida pelo cliente pudesse ser tratada de maneira devida, foram levantadas algumas hipóteses acerca das colunas e linhas existentes nesta. Seguem algumas destas hipóteses levantadas.

Hipótese 1: Churn por anos de serviço

Para uma análise efetiva dos casos em que o churn é mais provável, foi necessário correlacionar a taxa de churn com outros eventos. Um destes eventos é a quantidade de anos que a PowerCo prestou serviços ao cliente. Sendo assim, foi construído o gráfico de

barras apresentado nessa hipótese - utilizando a biblioteca matplotlib - o qual busca provar uma possível relação entre as colunas ‘churn’ e ‘num_years_antig’.

Gráfico 1 - Churn por anos de serviço



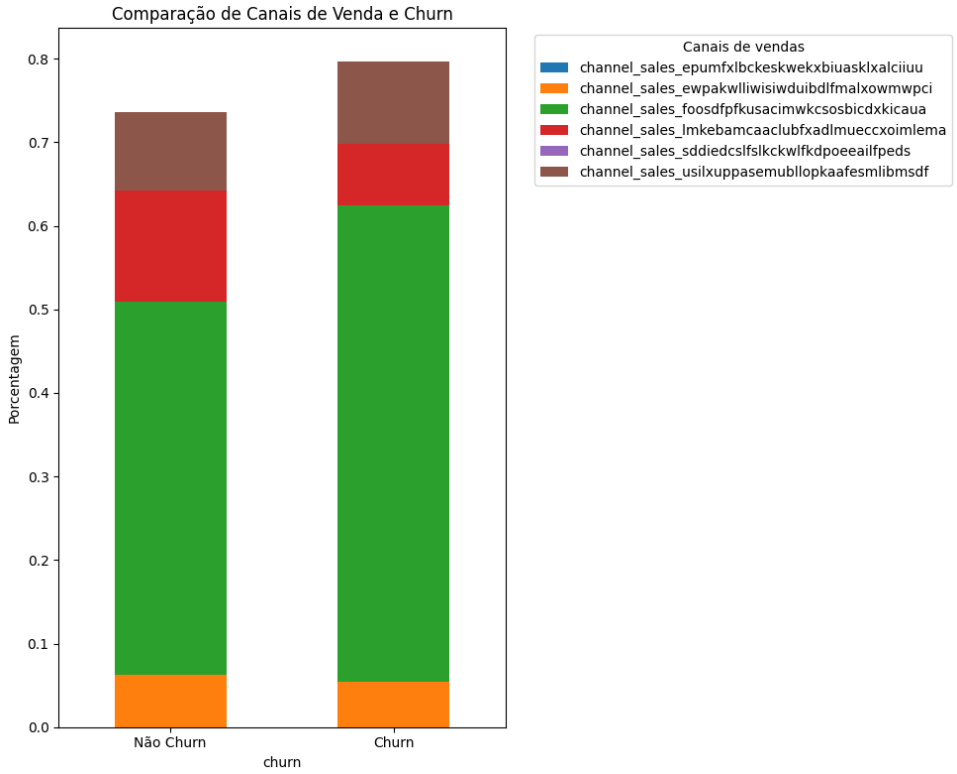
Fonte: Autoria Própria

Neste gráfico, somamos, às barras, os clientes de acordo com o tempo de serviço que eles têm na empresa. Depois disso, dividimos cada uma destas barras em clientes que deram churn e clientes que não deram churn. A partir disso, é perceptível que a faixa entre 3 e 6 anos é a faixa na qual se encontra o maior número de churns. Isto nos leva a inferir que na nossa solução devemos focar principalmente nos clientes que estão estão neste período tempo.

Hipótese 2: Canal de vendas

No banco de dados em que trabalhamos, há algumas colunas que possuem um número significativo de valores nulos. Uma destas é a ‘channel_sales’. No entanto, para que pudéssemos excluir esta coluna, precisamos verificar se esta interferia na taxa de churn ou não. Por isso, plotamos o seguinte gráfico:

Gráfico 2 - Relação entre churn e canal de vendas



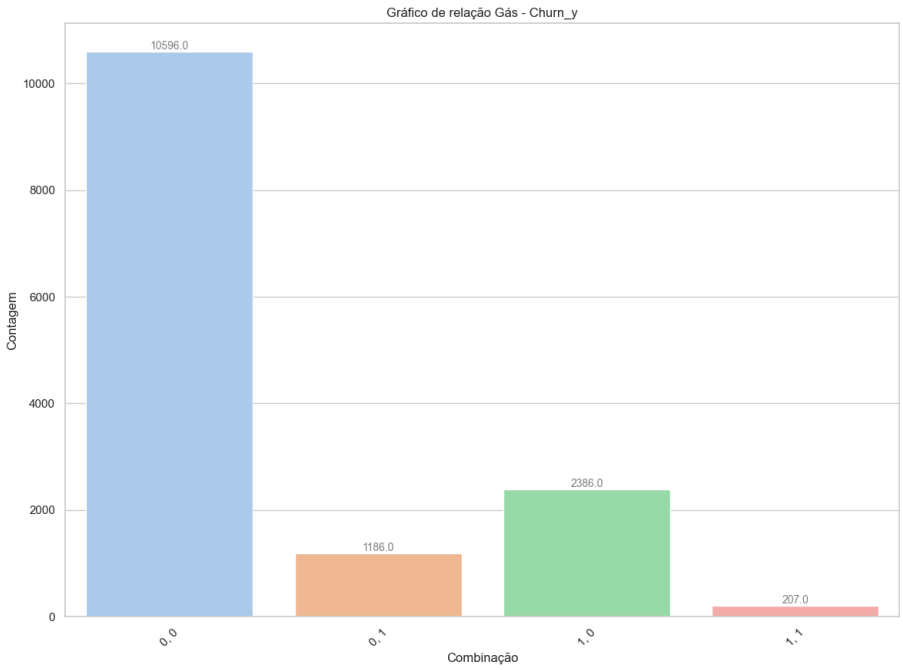
Fonte: Autoria Própria

Neste gráfico, podemos ver que o canal de vendas representado pela cor verde é o que possui uma maior taxa de churn. Por isso, acreditamos que mesmo que esta coluna possua muitos valores nulos, esta relação do canal de vendas com o churn pode ser valiosa para analisarmos a probabilidade de um cliente dar churn a partir do canal de vendas utilizado.

Hipótese 3: Churn por contrato de gás

Ao analisar o banco de dados pensamos que seria possível que dependendo do tipo de serviço utilizado pelo cliente poderia influenciar em sua probabilidade de churn, então pensando nisso plotamos um gráfico que representa a relação entre o uso do serviço de gás com o churn:

Gráfico 3 - Churn por contrato de gás



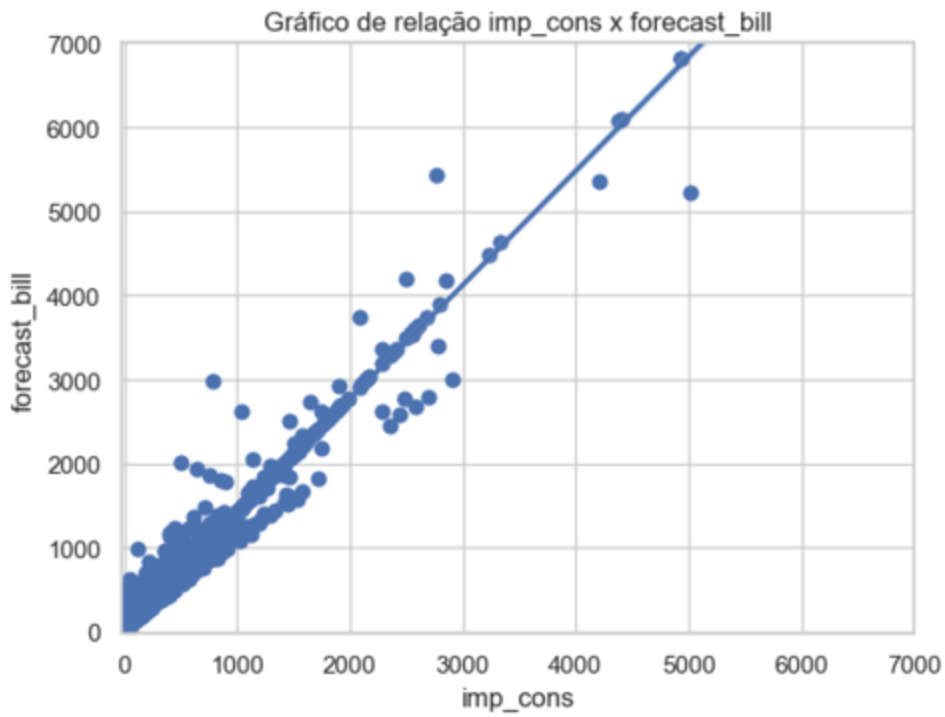
Fonte: Autoria própria

Como visto no gráfico, a probabilidade de um cliente realizar o churn sendo cliente de gás é proporcionalmente mais baixa do que no caso de não ser consumidor. Servindo assim como base para outras hipóteses e conclusões.

Hipótese 4: Relação de faturas

Após analisarmos as relações entre as colunas e seus comportamentos, fomos capazes de notar uma correlação entre as colunas que apontam as faturas pagas e as para serem pagas, gerando o seguinte gráfico:

Gráfico 4 - Relação de faturas



Fonte: Autoria própria

Levando em consideração os resultados que podem ser deduzidos desse gráfico que fizemos uma regressão linear para o preenchimento de uma coluna que apresentava dados inexistentes.

Logo, a partir das hipóteses anteriores, levantamos paradigmas para a construção do nosso modelo. No entanto, é válido afirmar que este é um processo iterativo. Assim, caso haja necessidade, poderemos retornar à construção de hipóteses e até mesmo anular alguma hipótese até o momento considerada válida.

4.3. Preparação dos Dados e Modelagem

A preparação de dados em machine learning envolve a coleta, limpeza e transformação de conjuntos de dados brutos em formatos adequados para treinar modelos de aprendizado de máquina. Isso inclui a remoção de dados ausentes, normalização e codificação de variáveis, entre outras tarefas. A modelagem refere-se à criação, treinamento e avaliação de diferentes algoritmos de machine learning. Esses processos são fundamentais para o sucesso de projetos de machine learning, garantindo a qualidade dos dados e a eficácia dos modelos.

À vista disso, a fim de tornar a modelagem do produto mais compreensível, cabe retomar seus objetivos descritos na seção 2.1: o modelo preditivo “Forecast” tem como objetivo prever quais clientes estão propensos a dar churn na companhia europeia de energia PowerCo. Para isso, o modelo deve ser de classificação e se basear numa base de

dados histórica da PowerCo, a qual contém, em suma, dados sobre os contratos firmados entre cliente e empresa, consumo de clientes, valores pagos por estes e valores despendidos pela instituição. Com tanto, o modelo se baseará nas informações potencialmente obtidas através de tais dados para definir se determinado cliente cometerá ou não churn, o que auxiliará o trabalho de gerentes de contas da empresa.

Portanto, foi testada a aplicação desse raciocínio em três tipos de algoritmos de aprendizagem de máquina diferentes — **XGBoost**, **Random Forest** e **SVM**. Passadas as etapas de exploração e pré-processamento de dados (discutidas, respectivamente, nas seções 4.2.1 e 4.2.2), foi possível dividir a base de dados em conjuntos de **treino** (para permitir a aprendizagem do algoritmo sobre os padrões e relações presentes na base) e **teste** (para validar o que foi aprendido pelo algoritmo) e inserí-los nos algoritmos mencionados, os quais terão seus funcionamentos e resultados discutidos de forma aprofundada mais adiante.

Contudo, a fim de compreender os resultados de forma adequada, é pertinente explicar as métricas utilizadas para medir a qualidade dos resultados obtidos. Logo, a seguir, constam as métricas utilizadas e descrições sucintas sobre elas:

Accuracy / Acurácia

Razão entre predições verdadeiras e predições no geral

Imagem 1 - Fórmula de Acurácia

$$A = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Fonte: Autoria própria

Precision / Precisão

Razão entre as predições positivas verdadeiras e as predições positivas no geral

Imagem 2 - Fórmula de Precisão

$$P = \frac{TP}{(TP + FP)}$$

Fonte: Autoria própria

Recall / Recuperação

Razão entre as predições positivas verdadeiras e a soma entre as predições positivas verdadeiras e as predições negativas falsas

Imagem 3 - Fórmula de Recall

$$R = \frac{TP}{(TP + FN)}$$

Fonte: Autoria própria

F1-Score

Razão entre o dobro do produto entre recall e precision e a soma entre recall e precision

Imagem 4 - Fórmula de F1-Score

$$F1 = \frac{2 \cdot P \cdot R}{(P + R)}$$

Fonte: Autoria própria

Matriz de confusão

Tabela que exibe e permite a comparação entre valores absolutos de predições:

- Positivas verdadeiras
- Positivas falsas
- Negativas verdadeiras
- Negativas falsas

Preparação de dados: escolha de features

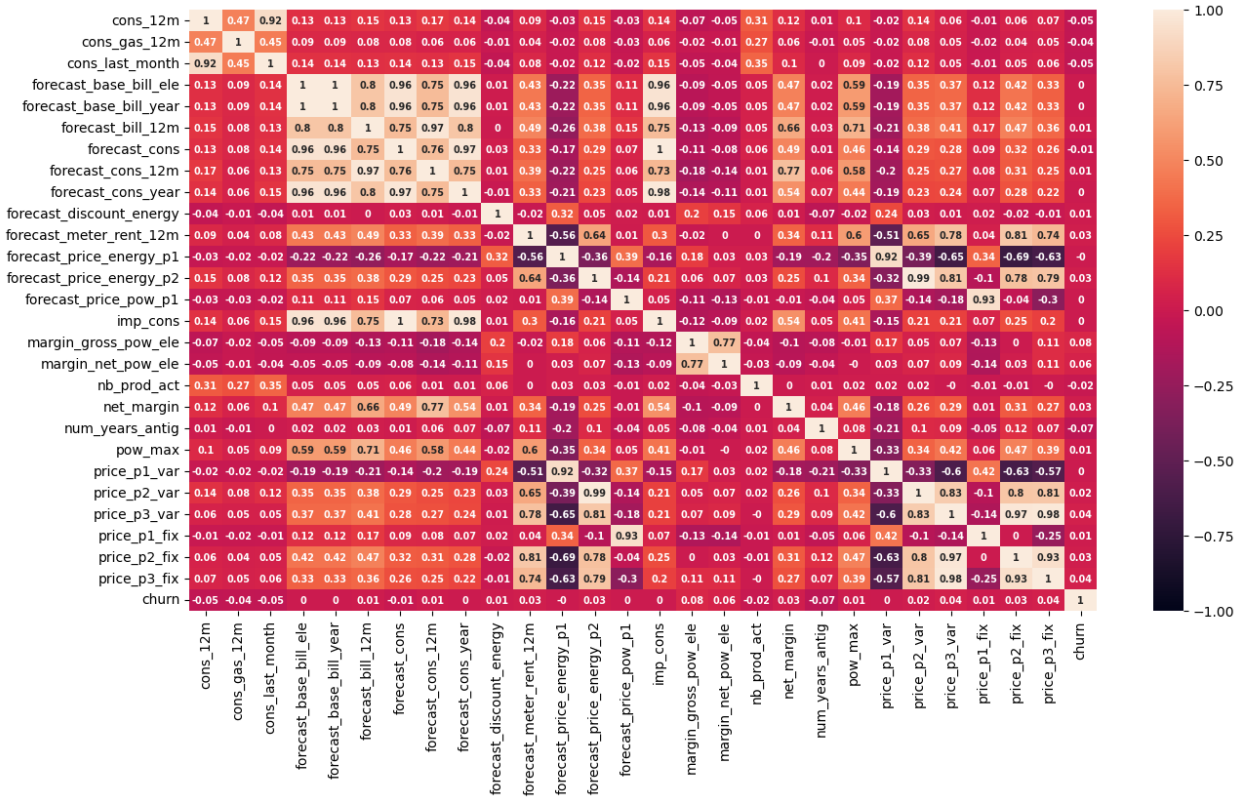
Havendo ciência da importância da preparação dos dados para uma boa modelagem, a equipe ForeSee trabalhou na seleção das melhores *features* — isto é, nesse contexto, colunas e seus respectivos dados — para serem utilizadas nos modelos preditivos. Para isso, com base na exploração de dados elaborada na seção 4.2.1, foi elaborada uma planilha contendo informações sobre cada uma das colunas presentes nas bases de dados do projeto, compreendendo desde a categorização dos dados até hipóteses sobre o impacto de cada coluna no churn (ou não) dos clientes. É possível acessar a planilha clicando aqui.

Contudo, as hipóteses formuladas dentro da planilha não possuem valor sem sustentação empírica. Portanto, também foi elaborado um mapa de calor com o valor de correlação entre cada uma das colunas de dados numéricos presentes na base de dados. Este mapa, presente na imagem a seguir, relaciona seus respectivos eixos (que contêm o título de cada coluna) por meio de células com valores que variam de -1 a 1. Quanto mais próximo de -1, maior é a tendência do valor de uma coluna aumentar ao passo que o valor da outra diminui; quanto mais próximo de 1, maior é a tendência dos valores de ambas as

colunas aumentarem ou diminuírem simultaneamente; no caso de valor próximo de 0, a tendência é que não exista correlação significativa entre as colunas em questão.

É importante ressaltar, no entanto, que, por conta da coluna *"churn"* ter caráter binário, é natural que os valores de correlação desta com as demais sejam próximos de 0. Todavia, isso não significa, necessariamente, que tais colunas não estejam correlacionadas.

Tabela 1 - Mapa de calor com valor de correlação



Fonte: Autoria própria

A partir dessas informações e de testes realizados com diversos tipos de modelos, foram inferidas diversas conclusões. Entre estas, cabe destaque para a questão do preço, que é intimamente ligado ao desconto, fator apontado como alavanca para impedir o churn de clientes. Diferente do que foi proposto como hipótese inicial pela PowerCo e pela própria equipe ForeSee na elaboração da planilha, apesar das colunas referentes a preço estarem correlacionadas com a de churn com valores relativamente próximos de 1, elas não se constituem como principais motivadoras ou desmotivadoras do churn, uma vez que outras colunas possuem valor de correlação com a coluna de churn na referida imagem de 0.05 ou abaixo de -0.05.

Entre tais colunas, cabe destaque para *"num_years_antig"*, referente ao número de anos ativos do cliente na PowerCo, *"margin_gross_pow_ele"* (referente à margem bruta na assinatura de energia) e "*margin_net_pow_ele"* (referente à margem líquida na assinatura de energia). Essas duas últimas, em especial, podem ser enfatizadas por terem sido apontadas,

na planilha, como hipoteticamente irrelevantes para o churn, o que foi visualmente negado a partir dessa segunda análise.

Ademais, também vale destacar que, analisando o mapa de calor, ao contrário do que foi apontado na planilha elaborada, nota-se que a maioria das colunas com valores previstos não têm grande relevância para o churn. Dentre estas, excedem-se *"forecast_meter_rent_12m"* e *"forecast_price_energy_p2"*, as quais ainda possuem coeficiente de correlação relativamente próximo de 0. Ainda assim, isso não significa que as colunas desse grupo, tais como outras, são totalmente dispensáveis para o modelo, pois ainda possuem um coeficiente mínimo de correlação e podem agregar nas previsões feitas pelo modelo.

Diante do exposto e de sucessivos testes com diferentes features e diferentes modelos, definiu-se como seleção de features ideal (até o momento) a seguinte:

Seleção de features

- id
- channel_sales
- cons_12m
- cons_gas_12m
- cons_last_month
- date_activ
- date_end
- date_modif_prod
- date_renewal
- forecast_base_bill_ele
- forecast_base_bill_year
- forecast_bill_12m
- forecast_cons_12m
- forecast_cons_year
- forecast_discount_energy
- forecast_meter_rent_12m
- forecast_price_energy_p1
- forecast_price_energy_p2
- forecast_price_pow_p1
- has_gas
- imp_cons
- margin_gross_pow_ele
- margin_net_pow_ele
- nb_prod_act
- net_margin
- num_years_antig
- origin_up

- pow_max
- price_p1_var
- price_p2_var
- price_p3_var
- price_p1_fix
- price_p2_fix
- price_p3_fix
- churn

Modelagem: resultados iniciais

A seguir, estão os 3 modelos candidatos (XGBoost, Random Forest e SVM) e a análise sobre seus primeiros resultados através das métricas de acurácia, precisão, recall e f1-score.

XGBoost:

Segundo a documentação do XGBoost, ele é um algoritmo de aprendizado de máquina amplamente utilizado para tarefas de classificação e regressão. Ele pertence à família de algoritmos de ensemble, que combinam várias árvores de decisão para obter um modelo mais poderoso e preciso. O XGBoost se destaca por sua eficiência e capacidade de lidar com conjuntos de dados complexos. Ele usa uma técnica chamada "gradient boosting" para treinar árvores sequencialmente, ajustando cada nova árvore para corrigir os erros das anteriores. Isso resulta em modelos robustos que podem capturar padrões sutis nos dados e evitar o overfitting.

```
  Acurácia de treino: 0.906695652173913

              precision    recall  f1-score   support

           0       0.91      1.00      0.95     10386
           1       1.00      0.04      0.07      1114

    accuracy                           0.91     11500
   macro avg       0.95      0.52      0.51     11500
weighted avg       0.92      0.91      0.87     11500


  Acurácia de teste: 0.9043478260869565

              precision    recall  f1-score   support

           0       0.90      1.00      0.95      2596
           1       1.00      0.01      0.03       279
```

```
    accuracy                           0.90      2875
   macro avg       0.95      0.51      0.49      2875
weighted avg       0.91      0.90      0.86      2875
```

Random Forest:

O Random Forest, de acordo com sua documentação, é outro algoritmo de ensemble que se baseia em árvores de decisão. Ele funciona criando várias árvores de decisão independentes durante o treinamento e, em seguida, combinando suas previsões para tomar uma decisão final. O "aleatório" no nome refere-se ao fato de que, durante o treinamento de cada árvore, amostras de dados e características são selecionadas aleatoriamente, o que ajuda a reduzir o overfitting e torna o modelo mais robusto. O Random Forest é conhecido por sua capacidade de lidar com conjuntos de dados grandes, com muitas características e classes, e é geralmente fácil de usar.

```
Acurácia de treino: 0.9999130434782608

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10386
           1       1.00      1.00      1.00      1114

    accuracy                           1.00     11500
   macro avg       1.00      1.00      1.00     11500
weighted avg       1.00      1.00      1.00     11500



Acurácia de teste: 0.9057391304347826

              precision    recall  f1-score   support

           0       0.91      1.00      0.95      2596
           1       1.00      0.03      0.06       279

    accuracy                           0.91      2875
   macro avg       0.95      0.51      0.50      2875
weighted avg       0.91      0.91      0.86      2875
```

SVM:

O SVM, conforme dito em sua documentação, é um algoritmo de aprendizado de máquina usado principalmente para classificação e regressão. Ele é eficaz na separação de classes em conjuntos de dados, encontrando um "hiperplano" que melhor divide as classes de maneira otimizada. O SVM é particularmente útil quando as classes são não linearmente separáveis, pois pode usar "kernels" para mapear os dados para um espaço de maior dimensão, onde a separação se torna possível. Isso significa que o SVM pode lidar com problemas complexos de classificação. Além disso, ele é eficiente em termos de uso de

memória, tornando-o uma escolha sólida para muitas aplicações de aprendizado de máquina.

Foram realizado testes utilizando 3 tipos de kernels:

- *Kernel Linear:* O kernel linear é o mais simples dos kernels. Ele realiza a separação de classes projetando os dados em um espaço de maior dimensão usando uma função linear. Isso significa que ele é eficaz quando as classes são linearmente separáveis, ou seja, podem ser divididas por uma linha reta ou hiperplano.
- *Kernel Radial Basis Function:* O kernel RBF é muito versátil e é frequentemente usado quando as classes não são linearmente separáveis. Ele mapeia os dados em um espaço de dimensão infinita usando funções de base radial. Isso permite que o SVM encontre fronteiras de decisão complexas e não lineares. O kernel RBF é útil para capturar padrões complexos nos dados.
- *Kernel Polinomial:* O kernel polinomial é semelhante ao kernel RBF, mas usa funções polinomiais em vez de funções de base radial. Isso permite que o SVM encontre fronteiras de decisão complexas e não lineares. O kernel polinomial é útil para capturar padrões complexos nos dados.

```
Acurácia de treino: 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10386
           1       1.00      1.00      1.00      1114

    accuracy                           1.00     11500
   macro avg       1.00      1.00      1.00     11500
weighted avg       1.00      1.00      1.00     11500



Acurácia de teste: 0.9050434782608696
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      2596
           1       0.57      0.08      0.14       279

    accuracy                           0.91      2875
   macro avg       0.74      0.54      0.55      2875
weighted avg       0.88      0.91      0.87      2875
```

Os 3 modelos testados performaram de maneira bastante similar, todos obtiveram uma acurácia no conjunto de testes próxima de 0.9 e todos exibiram taxas muito baixas de recall. Essas informações apontam para o um baixo número de predições verdadeiras para churn (**predição positiva verdadeira**) nos 3 casos. Ou seja, a taxa de recall permaneceu baixa porque houve muitos falsos negativos e a taxa de acurácia manteve-se elevada porque

a grande maioria do conjunto de testes é composto de negativos e o modelo previu um número muito maior de negativos do que de positivos.

Isolado esse problema comum, ainda é possível apontar diferenças nos resultados dos modelos e determinar qual se demonstra mais promissor para futuras análises. Ambos os modelos Random Forest e SVM apresentam um problema adicional nos resultados das métricas de avaliação: os resultados previstos no conjunto de treino atingiram 1.0 em todas as métricas, o que demonstra que os modelos estão em overfitting — ou seja, excessivamente adequados ao conjunto de treino. No caso do XGBoost esse problema não aparece e os resultados dos conjuntos de teste e treino, apesar de não serem iguais, são bastante próximos.

De forma mais palpável, os resultados iniciais verificados nestes testes indicam um ponto de preocupação sobre o impacto do Forecast na PowerCo. Sua atual inabilidade para predizer corretamente quais clientes darão churn (em detrimento da habilidade para predizer corretamente quem não dará churn) pode levar gerente de contas a tomarem decisões de negócios inadequadas, a exemplo de praticar medidas de retenção de cliente com um cliente erroneamente predito como tendente a churn. Tal situação pode gerar gastos desnecessários para a empresa e ir de desencontro com o objetivo do projeto, que é fundamentalmente, reduzir a perda financeira relacionada ao churn na PowerCo.

Pelos motivos citados, o modelo inicialmente escolhido como candidato é o XGBoost. Entretanto, este ainda deve ser modificado para que faça um maior número de predições positivas verdadeiras em prol do aumento da métrica de recall. Ademais, como o baixo recall foi um índice presente em todos os modelos testados, será feita uma reanálise do pré-processamento de dados em razão da possibilidade do problema ter origem nele.

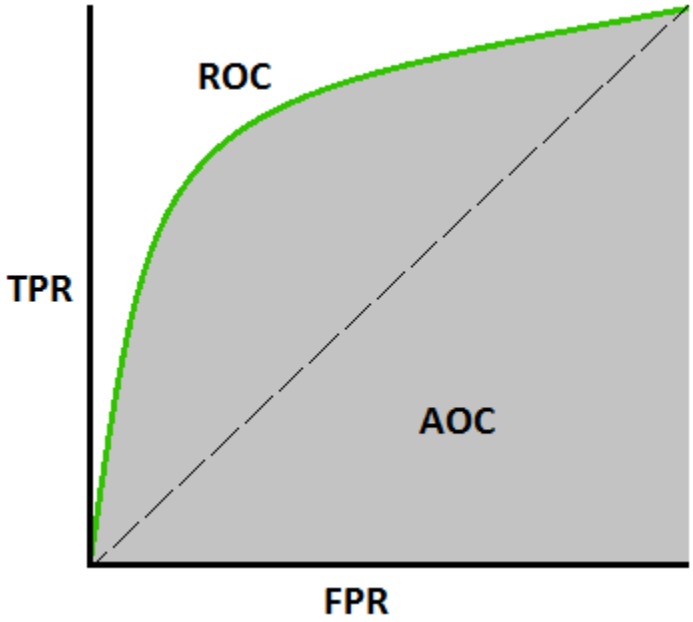
4.4. Comparação de Modelos

4.4.1 Discussão sobre as métricas

Retomando, uma métrica para um modelo preditivo é uma medida quantitativa que é usada para avaliar o desempenho e a precisão do modelo em fazer previsões sobre dados não vistos. Essas métricas são fundamentais para determinar o quão bem um modelo está se saindo em relação ao seu objetivo de previsão.

Tendo esta definição em pauta, cabe destacar que uma das principais métricas selecionadas para a avaliação é o **AUC-ROC** (*Area Under the Curve of Receiver Operating Characteristic* ou Área sob a Curva da Característica de Operação do Receptor), que é uma medida comum de desempenho para modelos de classificação binária. Ela se baseia num gráfico que relaciona, no eixo das abscissas, a taxa de falsos positivos e, no eixo das ordenadas, a taxa de verdadeiros positivos por meio de uma curva. O valor da métrica, em suma, representa a área abaixo dessa curva, sendo que, quanto maior for essa área, mais eficiente é o modelo. A imagem a seguir denota de forma visual esse gráfico:

Gráfico 5 - Exemplificação de AUC-ROC



Fonte: Towards Data Science - Medium

Além disso, também foi levada em consideração especial a métrica de recall. Isso se dá pela sua fórmula expressa na seção 4.3, que, assim como o AUC-ROC, prioriza as predições positivas verdadeiras e as compara com as predições negativas falsas. Em outras palavras, a escolha dessa métrica como um dos critérios principais de validação do modelo reduz a probabilidade de que um gerente de contas trate um cliente que dará churn como um cliente que não dará churn, o que significaria uma perda drástica de valor para a PowerCo, haja em vista a superioridade do custo de capturar novos clientes em comparação com o custo de manter clientes atuais.

Ademais, também foram usadas, como métrica, matrizes de confusão para cada um dos algoritmos testados para o Forecast. De forma concisa, elas exibem, lado a lado, os valores absolutos para a quantidade de cada tipo de previsão feita pelo modelo (**negativa verdadeira**, **positiva verdadeira**, **negativa falsa** e **positiva falsa**), o que permite uma visualização clara da forma pela qual o modelo está errando e/ou acertando suas previsões — e, também, uma comparação mais fácil entre modelos pela equipe ForeSee.

4.4.2 Acerca dos modelos eleitos

Dado que o objetivo final deste projeto é identificar quais clientes possuem uma maior probabilidade de dar churn, bem como medidas profiláticas para evitar tais ocorrências, tornou-se necessário eleger modelos que nos possibilitasse atingir o resultado desejado. Antes de explanar os modelos, é válido lembrar o conceito de modelo preditivo, que é um conjunto de algoritmos capaz de fazer previsões de eventos futuros a partir de padrões e

outras características identificadas em dados históricos. Além disso, também é crucial ressaltar que neste projeto é utilizado um modelo preditivo supervisionado, ou seja, para seu treinamento são utilizados dados que possuem rótulos. Desta forma, durante o treino, o algoritmo identifica padrões nos grupos de dados e observa a sua etiqueta. Assim, quando for testado, a partir dos padrões presentes nos dados, escolherá a etiqueta que mais se adequa a estes.

Durante o desenvolvimento do projeto, foram utilizados diversos modelos (algoritmos), dentre os quais pode-se dar mais destaque a três: **Random Forest**, **SVM-RBF** e **XGBoost**, os quais apresentaram os melhores desempenho após a aplicação da PCA (discutida na seção 4.2.2) e da tunagem de hiperparâmetros.

>> Tunagem de hiperparâmetros

É importante ressaltar que, nos modelos utilizados, para a definição de hiperparâmetros, foi utilizada a Random Search. Esta técnica busca encontrar valores ótimos para hiperparâmetros seguindo alguns passos para utilizá-la. O primeiro deles é especificar o intervalo dentro do qual se buscará os hiperparâmetros. Após isto, faz-se necessário definir o número de iterações que se deseja realizar, sabendo que em cada uma destas ocorrerá a seleção aleatória de hiperparâmetros dentro do intervalo já especificado. Para cada um destes conjuntos de hiperparâmetros, executa-se o modelo, avaliando sempre as métricas definidas e entendendo qual conjunto de hiperparâmetros desempenhou melhor segundo estas.

No entanto, outra alternativa seria a utilização da Grid Search, técnica que também visa a otimização de hiperparâmetros para modelos de aprendizado de máquina. A aplicação desta técnica consiste em definir um intervalo de busca para os hiperparâmetros que se deseja otimizar, após isto, o algoritmo cria e testa todas as combinações possíveis de hiperparâmetros dentro do intervalo previamente definido. Após isso, é necessário treinar e avaliar o modelo para cada combinação na grade, isto a partir da métrica que mais se adequa ao caso.

Desta forma, é importante lembrar que estas técnicas diferem pois, enquanto na Random Search são selecionados valores aleatórios para cada iteração envolvendo o conjunto de hiperparâmetros, na Grid Search, todas as combinações possíveis de hiperparâmetros considerados são testadas. Isto faz com que, no caso de desenvolvimento do Forecast, aquela seja uma técnica mais eficiente, por não exigir um processamento computacional muito alto, enquanto esta, por testar todas as combinações, seja uma operação de alto custo computacional e, por isso, mais demorada e cara. Além disso, a Random Search é mais indicada quando não se há uma noção clara do impacto dos hiperparâmetros no modelo, enquanto a Grid Search é recomendada para casos onde se há uma compreensão sólida do impacto e da relação dos hiperparâmetros. Por estes motivos, neste projeto, optou-se por dar prioridade à utilização da Random Search, dado que o grupo não dispunha de alto poder computacional, bem como não possuía a certeza do impacto de hiperparâmetros no modelo.

>> Análise e comparação dos modelos testados

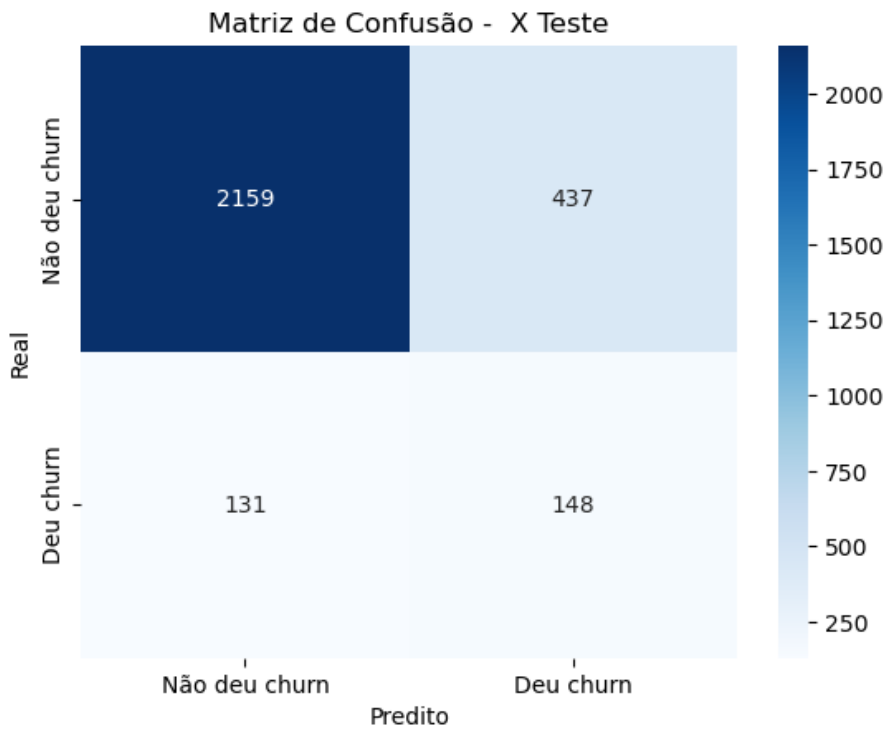
> XGBoost

Retomando a definição de XGBoost da seção 4.3:

> XGBoost, segundo sua própria documentação, é um algoritmo de aprendizado de máquina amplamente utilizado para tarefas de classificação e regressão. Ele pertence à família de algoritmos de ensemble, que combinam várias árvores de decisão para obter um modelo mais poderoso e preciso. O XGBoost se destaca por sua eficiência e capacidade de lidar com conjuntos de dados complexos. Ele usa uma técnica chamada "gradient boosting" para treinar árvores sequencialmente, ajustando cada nova árvore para corrigir os erros das anteriores. Isso resulta em modelos robustos que podem capturar padrões sutis nos dados e evitar o overfitting.

Após a aplicação do PCA e a tunagem de hiperparâmetros, o XGBoost obteve melhoras significativas em sua predições na base de dados de treino. Isso pode ser visto por meio das métricas de AUC-ROC e Recall que foram atingidas através do algoritmo, respectivamente: **0.68** e **0.53**. A melhoria nos resultados torna-se ainda mais evidente ao comparar as 148 previsões positivas verdadeiras com as 131 previsões negativas falsas na matriz de confusão presente na tabela 2, gerada a partir da aplicação do algoritmo:

Tabela 2 - Matriz de confusão: XGBoost



Fonte: Autoria própria

Para alcançar esses resultados, a tunagem de hiperparâmetros via Random Search considerou os seguintes hiperparâmetros:

Hiperparâmetros tunados no XGBoost

-> n_estimators (Número de Estimadores)

- Este hiperparâmetro determina o número de árvores de decisão que serão construídas durante o treinamento do modelo. Quanto maior o número de estimadores, mais complexo o modelo será. Um valor maior geralmente melhora o desempenho do modelo, mas também aumenta o tempo de treinamento. No entanto, é importante evitar um número excessivamente alto, pois pode levar ao overfitting. [valor selecionado: 450]

-> max_depth (Profundidade Máxima da Árvore)

- Define a profundidade máxima de cada árvore de decisão no modelo. Uma árvore mais profunda pode representar relacionamentos mais complexos nos dados, mas também pode levar ao overfitting. Um valor maior aumenta a complexidade do modelo, mas também o torna mais propenso a overfitting. Encontrar a profundidade adequada é uma parte crítica do ajuste de hiperparâmetros. [valor selecionado: 8]

-> learning_rate (Taxa de Aprendizado)

- Este hiperparâmetro controla a taxa na qual o modelo aprende com os erros anteriores. Um valor menor torna o aprendizado mais lento, enquanto um valor maior permite um aprendizado mais rápido. Valores menores resultam em modelos mais robustos, mas podem exigir um número maior de estimadores para atingir o mesmo desempenho que uma taxa de aprendizado maior. É importante encontrar um equilíbrio. [valor selecionado: 0.1]

-> subsample (Subamostragem)

- Especifica a fração de observações (amostras) que são usadas para treinar cada árvore. Um valor menor cria subamostras aleatórias, o que pode ajudar a evitar overfitting. Usar uma fração menor pode melhorar a generalização do modelo e reduzir o overfitting, mas também pode fazer com que o modelo seja menos preciso. [valor selecionado: 0.8]

-> colsample_bytree (Subamostragem de Colunas por Árvore)

- Define a fração de recursos (colunas) que são amostrados aleatoriamente para construir cada árvore. Isso introduz aleatoriedade na construção de cada árvore. A subamostragem de colunas pode ajudar a evitar a correlação entre as árvores e melhorar a generalização do modelo. Valores menores introduzem mais aleatoriedade. [valor selecionado: 1.0]

-> gamma (Poda Mínima)

- O parâmetro de poda mínima (também conhecido como “minimum loss reduction”) controla quando uma árvore será dividida com base em uma métrica de ganho. Um valor maior requer ganhos maiores para dividir um nó, tornando o modelo mais conservador. Um valor maior de gamma torna o modelo mais conservador, evitando a formação de árvores muito profundas e complexas. [valor selecionado: 0]

-> min_child_weight (Peso Mínimo da Folha)

- Este hiperparâmetro especifica a soma mínima dos pesos (ou instâncias) em cada folha da árvore. Isso pode ser usado para evitar divisões em folhas com poucas amostras. Valores maiores tornam o modelo mais conservador, evitando folhas com poucas amostras, o que pode ajudar a evitar overfitting. [valor selecionado: 1]

> Random Forest

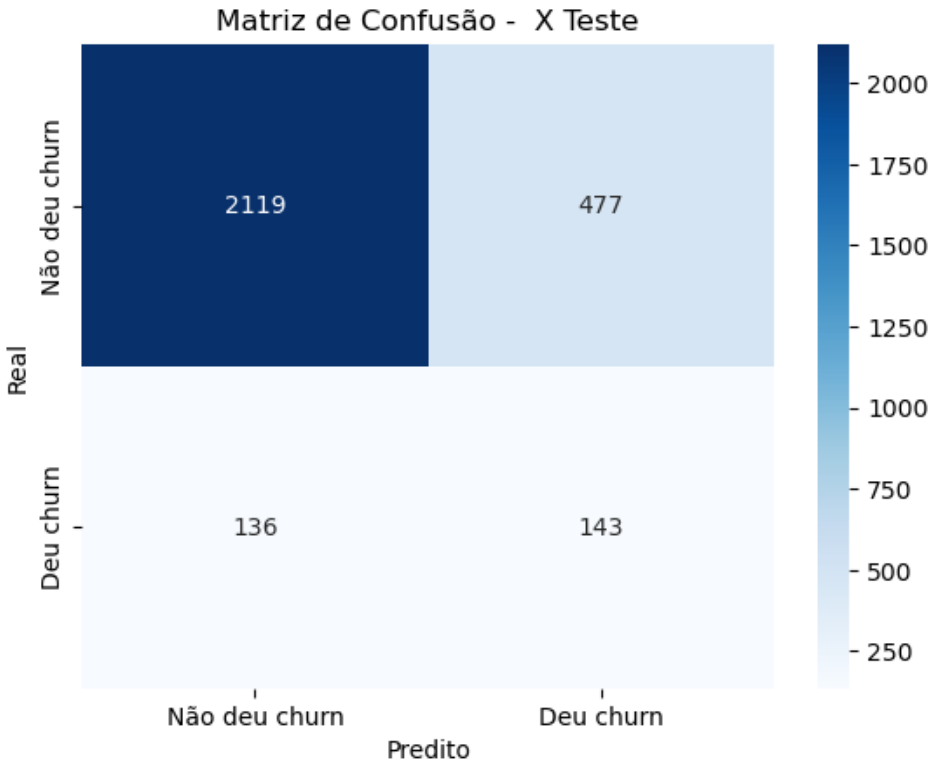
Retomando a definição de Random Forest da seção 4.3:

O Random Forest, de acordo com sua documentação, é outro algoritmo de ensemble que se baseia em árvores de decisão. Ele funciona criando várias

árvores de decisão independentes durante o treinamento e, em seguida, combinando suas previsões para tomar uma decisão final. O “aleatório” no nome refere-se ao fato de que, durante o treinamento de cada árvore, amostras de dados e características são selecionadas aleatoriamente, o que ajuda a reduzir o overfitting e torna o modelo mais robusto. O Random Forest é conhecido por sua capacidade de lidar com conjuntos de dados grandes, com muitas características e classes, e é geralmente fácil de usar.

Com a otimização da implementação do modelo em Random Forest por meio da aplicação do PCA e da tunagem de hiperparâmetros, foram alcançados os respectivos valores para AUC-ROC e Recall: **0.66** e **0.51**. Há, ainda, a matriz de confusão da tabela 3, que mostra a quantidade de cada tipo de predição realizada com a implementação desse algoritmo na base de dados.

Tabela 3 - Matriz de confusão: Random Forest



Fonte: Autoria própria

Como se observa, os quadrantes inferiores da matriz de risco possuem valores bem próximos. Isso significa que, através do Random Forest, o modelo não é capaz de definir, com eficiência, se uma pessoa que praticou churn realmente o praticou ou não. Essa informação, junto aos valores das métricas citadas anteriormente, demonstra que, apesar da melhoria em relação à aplicação pré-PCA e pré-tunagem do Random Forest, tal algoritmo ainda não supera o desempenho do XGBoost.

Os resultados em questão foram alcançados por meio do ajuste via Random Search dos seguintes hiperparâmetros:

Hiperparâmetros tunados no Random Forest

-> n_estimators (Número de Estimadores)

- Especifica o número de árvores de decisão independentes que serão construídas no modelo Random Forest. Cada árvore contribui com uma votação para as previsões finais. Aumentar o número de estimadores geralmente melhora o desempenho do modelo, mas também aumenta o tempo de treinamento. No entanto, é importante evitar valores excessivamente altos para evitar o overfitting. [valor selecionado: 120]

-> max_depth (Profundidade Máxima da Árvore)

- Define a profundidade máxima permitida para cada árvore de decisão no Random Forest. Árvores mais profundas podem representar relações mais complexas nos dados, mas também aumentam o risco de overfitting. Um valor maior aumenta a complexidade do modelo, mas pode torná-lo mais propenso ao overfitting. Encontrar a profundidade adequada é importante para o ajuste de hiperparâmetros. [valor selecionado: 60]

-> bootstrap (Amostragem com Substituição)

- Este hiperparâmetro controla se a amostragem é realizada com ou sem substituição. Quando definido como True, cada árvore é treinada em uma amostra bootstrap aleatória dos dados de treinamento, o que introduz variação nas amostras. A amostragem com substituição ajuda a introduzir variação nos modelos individuais e melhora a generalização do Random Forest. [valor selecionado: False]

-> class_weight (Peso das Classes)

- Este hiperparâmetro permite atribuir pesos diferentes às classes de destino no caso de um conjunto de dados desequilibrado. Os pesos ajudam a dar mais importância às classes minoritárias. Usar pesos de classe adequados é crucial em conjuntos de dados desequilibrados, pois ajuda o modelo a dar uma atenção adequada às classes minoritárias. [valor selecionado: ‘balanced’]

> SVM

Retomando a definição de SVM da seção 4.3:

> O SVM, conforme dito em sua documentação, é um algoritmo de aprendizado de máquina usado principalmente para classificação e regressão. Ele é eficaz na separação de classes em conjuntos de dados, encontrando um "hiperplano" que melhor divide as classes de maneira otimizada. O SVM é particularmente útil quando as classes são não linearmente separáveis, pois pode usar "kernels" para mapear os dados para um espaço de maior dimensão, onde a separação se torna possível. Isso significa que o SVM pode lidar com problemas complexos de classificação. Além disso, ele é eficiente em termos de uso de memória, tornando-o uma escolha sólida para muitas aplicações de aprendizado de máquina.

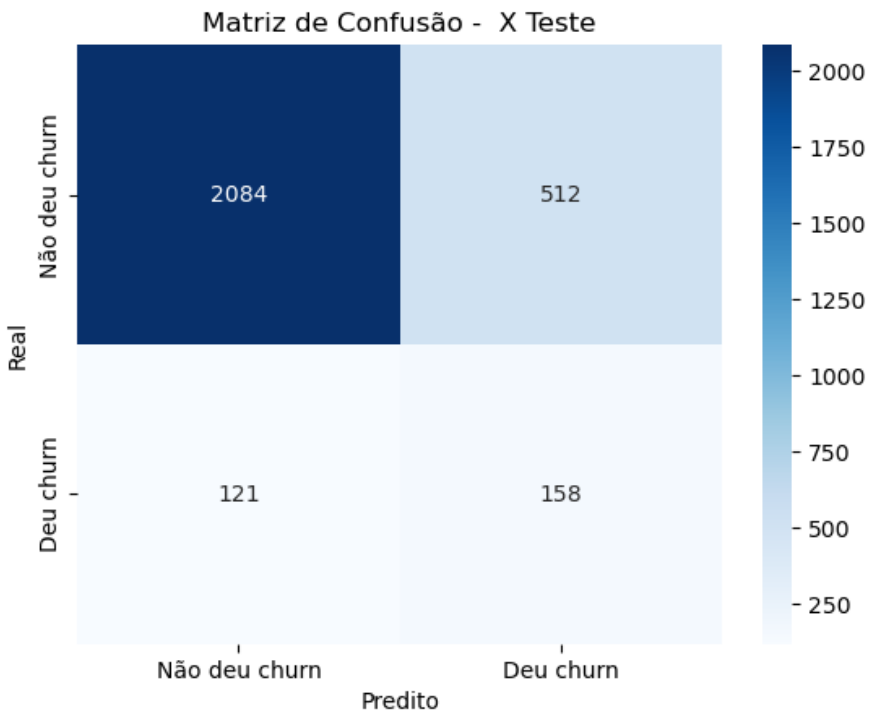
À vista disso, vale ressaltar que o Kernel escolhido para a modelagem por SVM foi o Radial Basis Function (ou RBF), cuja definição a seguir foi retomada, também, da seção 4.3:

> O kernel RBF é muito versátil e é frequentemente usado quando as classes não são linearmente separáveis. Ele mapeia os dados em um espaço de dimensão infinita usando funções de base radial. Isso permite que o SVM encontre fronteiras de decisão complexas e não lineares. O kernel RBF é útil para capturar padrões complexos nos dados.

A partir desse contexto, a otimização do SVM também se deu sob a aplicação do PCA e da tunagem de hiperparâmetros, mas sem o uso de Random Search ou Grid Search. Em vez disso, fez-se um ajuste manual de hiperparâmetros, haja vista que as duas demais técnicas supracitadas exigiram, para o conjunto limitante de parâmetros selecionados pela ForeSee, um poder de processamento computacional indisponível para a equipe.

Dessa forma, as métricas de AUC-ROC e Recall para o algoritmo de SVM apresentaram valores respectivos de **0.68** e **0.57**. Esses resultados, juntos à matriz de confusão da tabela 4, denotam que, apesar da dificuldade enfrentada para ajustar os hiperparâmetros do SVM de maneira ideal, este algoritmo teve desempenho semelhante ao XGBoost, sendo capaz, até, de classificar mais corretamente os clientes da PowerCo que praticaram churn.

Tabela 4 - Matriz de confusão: SVM



Fonte: Autoria própria

Esses resultados foram possíveis através do ajuste dos seguintes hiperparâmetros:

Hiperparâmetros tunados no SVM

-> C (Parâmetro de Regularização)

- Este hiperparâmetro controla a força da regularização. Um valor menor especifica uma regularização mais forte, o que significa que o modelo tentará ajustar os dados de treinamento o máximo possível. Um valor maior especifica uma regularização mais fraca, o que significa que o modelo tentará, além dos dados de treinamento, também aprender padrões mais complexos. É importante encontrar um equilíbrio. [valor selecionado: 0.1]

-> kernel (Função Kernel)

- Este hiperparâmetro especifica o tipo de função kernel a ser usado pelo modelo. O kernel é usado para mapear os dados para um espaço de maior dimensão, onde a separação se torna possível. O kernel linear é o mais simples e é usado quando as classes são linearmente separáveis. O kernel RBF é muito versátil e é usado quando as classes não são linearmente separáveis, tal qual o polinomial. [valor selecionado: ‘rbf’]

-> gamma (Coeficiente do Kernel)

- Este hiperparâmetro é usado apenas para os kernels RBF e polinomial. Ele define o quão longe a influência de um único exemplo de treinamento alcança, com valores baixos significando 'longe' e valores altos significando 'perto'. Valores baixos significam que as instâncias de treinamento têm uma influência mais distante, o que significa que o hiperplano de decisão é mais suave. Valores altos significam que as instâncias de treinamento têm uma influência mais próxima, o que significa que o hiperplano de decisão é mais ajustado. [valor selecionado: 'scale']

-> shrinking (Encolhimento)

- Este hiperparâmetro controla se o modelo usará ou não a heurística de encolhimento. A heurística de encolhimento é uma técnica que acelera o treinamento do modelo, reduzindo o número de vetores de suporte. Isso pode melhorar o desempenho do modelo, mas também pode reduzir a precisão. [valor selecionado: True]

-> class_weight (Peso das Classes)

- Este hiperparâmetro permite atribuir pesos diferentes às classes de destino no caso de um conjunto de dados desequilibrado. Os pesos ajudam a dar mais importância às classes minoritárias. Usar pesos de classe adequados é crucial em conjuntos de dados desequilibrados, pois ajuda o modelo a dar uma atenção adequada às classes com menor número de amostras. [valor selecionado: 'balanced']

-> probability (Probabilidade)

- Este hiperparâmetro controla se o modelo deve ou não calcular probabilidades para as previsões. Isso pode ser útil para alguns algoritmos, como o SVM, que não são capazes de calcular probabilidades por padrão. No entanto, isso pode aumentar o tempo de treinamento. Desse modo, é possível analisar o churn através de um threshold. [valor selecionado: True]

4.4.3 Definição do modelo escolhido

A seleção do modelo preditivo final é uma etapa crítica em qualquer projeto de aprendizado de máquina e, após uma análise rigorosa, o XGBoost emerge como a escolha mais sólida e confiável para este projeto. Sua superioridade em termos de desempenho, capacidade de lidar com dados complexos, resistência ao overfitting e flexibilidade de customização destacam o XGBoost como a ferramenta ideal para enfrentar os desafios específicos apresentados por este contexto. Além dos valores de AUC-ROC e Recall serem respectivamente 0.68 e 0.53, valores excelentes. Em adendo, os seus erros são de certa forma "positivos" por apresentar relativamente poucos erros de falsos não churns

Além disso, a escalabilidade do XGBoost é um fator crucial para garantir que o modelo possa lidar de forma eficaz com o crescimento e a diversificação dos dados ao longo do tempo. Isso proporciona uma base sólida para a aplicabilidade contínua do modelo à medida que o projeto evolui. Por outro lado, as outras opções de modelos explorados na seção 4.4.2 não apresentavam tamanhas vantagens em relação a suas especificidades e métricas, tornando o XGBoost a escolha clara para o projeto.

4.5. Avaliação

Em face do processo de desenvolvimento descrito ao longo deste documento, a solução final proposta sob o nome de **Forecast** se consolidou na sua forma de MVP (*Minimum Viable Product* / Produto Minimamente Viável). Portanto, de forma adequada à proposta inicial do projeto e ao entendimento do negócio discorrido durante a seção 4.1, tal versão conta com um arquivo principal do tipo Jupyter Notebook que apresenta, como partes essenciais, uma etapa de pré-processamento de dados, um modelo preditivo baseado no algoritmo XGBoost e um sistema de recomendação de desconto para conter o churn de clientes da PowerCo.

À vista disso, cabe retomar ou introduzir a explicação do funcionamento de cada uma dessas partes. Logo, como vê-se detalhadamente na seção 4.2.2, em relação ao pré-processamento de dados, foi feita a deleção de linhas e colunas com excesso de dados nulos, o preenchimento de colunas referentes a previsão de preço através de regressão linear, a correção de dados ruidosos (como datas impossíveis e valores não numéricos para colunas binárias), a normalização dos valores numéricos, a aplicação de PCA e o balanceamento da base de dados através de SMOTE. Todo esse processo é feito com o objetivo de possibilitar a leitura adequada dos dados pelo modelo preditivo e, consequentemente, a execução ideal de suas predições.

Feito o pré-processamento, dá-se início à execução do modelo preditivo em si. Com base no processo de modelagem de dados descrito ao longo da seção 4.3 e, também, na seção 4.4, verificou-se que o melhor algoritmo de aprendizagem de máquina para o desenvolvimento do Forecast como modelo preditivo de classificação é o XGBoost. Esse algoritmo trabalha com uma aplicação extrema — isto é, com um grande número de iterações e com uma grande variação nos dados em cada iteração — da técnica de *"Gradient Boosting"*, que combina vários algoritmos pouco eficientes para, com a autocorreção destes a cada iteração, formar um algoritmo adequadamente eficaz.

Diante disso, os dados já pré-processados são analisados pelo modelo e este, por sua vez, fornece as predições para cada cliente cujos dados foram analisados, afirmando se este cometerá churn ou não. Com essas predições, a execução do sistema de recomendação, elaborado como ferramenta para definir a taxa de desconto ideal para prevenir o churn de cada cliente, torna-se possível.

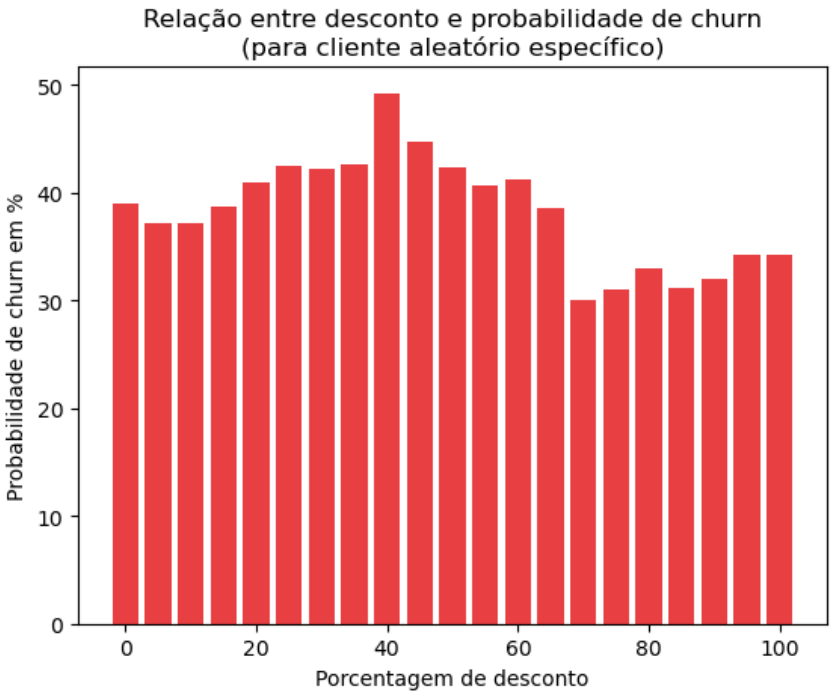
Sistema de recomendação

O sistema de recomendação elaborado pela equipe ForeSee se trata de um algoritmo iterativo sobre os dados inseridos no modelo pelo usuário, conforme as seguintes etapas:

1. O modelo preditivo realiza as predições sobre a base de dados inserida no notebook pelo usuário;
2. O sistema de recomendação recebe todas as linhas de dados para os quais o modelo preditivo previu “churn”;
3. A partir de cada uma dessas linhas, o sistema de recomendação cria novas linhas aplicando diferentes valores de desconto para a coluna *“forecast_discount_energy”*, variando tal desconto de 0% a 50%;
4. Essas novas linhas são reanalisadas pelo modelo preditivo, que retornará a probabilidade do cliente não praticar churn para cada uma delas;
5. Com isso, o sistema de recomendação irá calcular o EV (Expected Value / Valor Esperado) de cada linha multiplicando o valor de tal probabilidade pelo valor de margem líquida total na coluna *“net_margin”*;
6. Por fim, para cada cliente cuja predição original apontava para churn, o sistema de recomendação retornará a linha que proporcionar maior EV.

A partir da construção e da testagem do sistema de recomendação, foi possível complementar a hipótese inicial trazida pela PowerCo, referente à ideia de que o desconto sobre o preço pago pelos clientes da empresa impacta na chance dos clientes praticarem churn. Essa hipótese, inicialmente, foi verificada pela equipe ForeSee por meio da submissão de dados artificiais idênticos (variando apenas a porcentagem de desconto) para o Forecast, que atribuiu diferentes valores de probabilidade de churn para cada um deles. A partir disso, a construção do seguinte gráfico foi possível:

Gráfico 6 - Relação entre desconto e probabilidade de churn (para cliente aleatório específico)



Fonte: Autoria própria

De acordo com o gráfico 6 e com o processo para sua construção, é nítido que, como reclama a **hipótese inicial**, agora acatada, da PowerCo, a porcentagem de desconto tem impacto sobre a probabilidade do cliente praticar churn. Contudo, não se trata de uma relação simples de proporcionalidade inversa: a probabilidade do cliente praticar churn sob um determinado valor de desconto pode ser maior do que a mesma probabilidade sob um valor de desconto menor. Logo, o sistema de recomendação se consolida como uma ferramenta de auxílio vital para o usuário do Forecast, uma vez que permite-o definir a porcentagem de desconto mais financeiramente vantajosa para a PowerCo de acordo com os dados de cada cliente analisado.

Ainda assim, é indispensável salientar que tanto o modelo preditivo quanto o sistema de recomendação não são isentos de erro. Por isso, faz-se essencial a formulação de um **plano de contingência** referente à erros oriundos do manuseio do produto.

O plano de contingência para o modelo preditivo é concebido através de uma abordagem estratégica que se baseia no cálculo do impacto financeiro direto sobre o lucro da empresa em relação aos fundos potencialmente perdidos devido a falhas no modelo. Esta abordagem visa quantificar e qualificar os riscos associados à imprecisão das previsões através da multiplicação do desconto pelo preço que seria pago para todos os falsos positivos, permitindo à empresa tomar decisões informadas sobre a implementação de medidas corretivas. Ao considerar não apenas o custo imediato da falha, mas também o efeito cascata sobre a rentabilidade, este plano proporciona uma visão abrangente da

importância crítica da precisão do modelo preditivo no contexto do desempenho financeiro da organização.

5. Conclusões e Recomendações

Em vista da conclusão do processo de desenvolvimento de MVP do Forecast, é viável e proveitoso realizar uma avaliação geral sobre o produto alcançado. Desse modo, é possível reafirmar que o Forecast se consolida como um modelo preditivo baseado no algoritmo de classificação XGBoost para prever quais clientes tendem ao churn na PowerCo.

Ao decorrer das etapas de desenvolvimento e otimização do modelo, foram alcançados resultados satisfatórios. Em termos de métricas, o Forecast conta com 69% de AUC-ROC, 95% de precisão e 83% de recall, valores que podem aumentar com futuros aprimoramentos por cientistas de dados da PowerCo, haja em vista que a escalabilidade do XGBoost foi um dos motivos para sua escolha como algoritmo de aprendizagem de máquina escolhido para o produto.

Entretanto, mesmo com métricas satisfatórias, tanto o modelo preditivo quanto o sistema de recomendação incluso como parte do Forecast são passíveis de erro em suas predições. Logo, os gerentes de contas e demais usuários do produto devem, além de seguir adequadamente as instruções de usos que constam em vídeo anexado no arquivo README.md do repositório do projeto, se atentar a predições que se demonstrarem incoerente e se adequar ao modelo de plano de contingência proposto na seção 4.5.

Ademais, também é proveitoso para a PowerCo — e para a público afetado pelo modelo no geral — que os usuários do produto se orientem pelas recomendações formais e as considerações éticas a seguir:

Recomendações Formais

1. Implementação Cautelosa:
- Realizar a implementação do modelo em um ambiente controlado e monitorado para validar as previsões em um cenário real e ajustar qualquer parâmetro, se necessário.
2. Monitoramento Contínuo:
- Estabelecer um protocolo de monitoramento contínuo para o modelo a fim de garantir sua validade e eficácia a longo prazo, identificando quaisquer desvios ou alterações nas tendências de churn.
3. Aprimoramento Periódico:
- Implementar ciclos regulares de revisão e recalibração do modelo, assegurando que este continue a ser alimentado com dados atuais e relevantes para manter sua precisão e utilidade.
4. Treinamento e Suporte:
- A equipe de análise de dados deve passar por treinamento adequado para a utilização e interpretação do modelo e suas previsões.

Considerações Éticas

1. Comunicação Clara:
- Assegure-se de que a comunicação com os clientes seja transparente, especialmente ao implementar estratégias para reter aqueles identificados como propensos ao churn.
2. Proteção de Dados:
- Garanta que todas as informações pessoais utilizadas e/ou geradas pelo modelo sejam armazenadas, processadas e protegidas em conformidade com as leis e regulamentações locais de proteção de dados.

Com o devido atendimento aos tópicos supracitados, espera-se que o Forecast empenhe seu papel da maneira mais eficiente possível.

Conclusão

A implementação estratégica do modelo Forecast não somente possibilitará a identificação de clientes em potencial risco de churn, mas também habilitará a PowerCo a desenvolver e implementar estratégias de retenção mais eficazes. A manutenção e a operacionalização ética deste modelo são fundamentais para sustentar uma relação de confiança com os clientes e garantir a conformidade reputacional da empresa no mercado.

6. Referências

AFFONSO, A. Você Sabe O Que É a Metodologia CRISP DM? Disponível em: https://www.voitto.com.br/blog/artigo/crisp-dm.

The new E.ON: intelligent power grids and customer solutions. Disponível em: https://www.eon.com/en.html

E.ON. Our privacy policy | Using your personal data | E.ON. Disponível em: https://www.eonenergy.com/privacy.html.

ENEL Green Power | A transição energética na Europa e na América Latina: em que ponto estamos? Disponível em: https://www.enelgreenpower.com/pt/learning-hub/debates/transicao-energetica-europa-america-latina.

ÉPOCA Negócios | Gigantes de energia da Europa anunciam aliança. Disponível em: https://epocanegocios.globo.com/sustentabilidade/noticia/2023/05/gigantes-de-energia-da-europa-anunciam-alianca.ghtml.

Universidade Federal da Integração Latino-Americana | Matriz de Risco. Disponível em: https://portal.unila.edu.br/proagi/cccl/demandantes-e-area-tecnica/matriz-de-risco. Acesso em: 10 set. 2023.

Kellison Ferreira | Canvas de proposta de valor: para que serve e como preencher. Disponível em: https://blog.somostera.com/product-management/canvas-de-proposta-de-valor. Acesso em: 11 set. 2023.

xgboost developers | xgboost 1.7.2 documentation. Disponível em: https://xgboost.readthedocs.io/en/stable/#.

SCIKIT-LEARN. sklearn.ensemble.RandomForestClassifier | scikit-learn 0.20.3 documentation. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

SCIKIT LEARN. 1.4. Support Vector Machines | scikit-learn 0.20.3 documentation. Disponível em: https://scikit-learn.org/stable/modules/svm.html.

Persona Rikke Friis Dam and Teo Yu Siang | A Simple Introduction | Interaction Design Foundation. Disponível em: https://www.interaction-design.org/literature/article/personas-why-and-how-you-should-use-them. Acesso em: 11 set. 2023.

Portal Sebrae - Sebrae | Ferramenta: 5 FORÇAS DE PORTER (CLÁSSICO). Disponível em: https://sebrae.com.br/Sebrae/Portal%20Sebrae/Anexos/ME_5-Forcas-Porter.PDF. Acesso em: 12 out. 2023.

Anexos