Documentação da Manipulação de dados

Para treinar nosso modelo, nossa equipe primeiramente processou os dados em duas etapas, Pré-processamento e Feature engineering.

Etapas:

1- Pré-processamento: Nosso foco nessa etapa é acabar com os dados faltantes. Em cada coluna com faltantes: imputamos valores, dropamos as linhas de nulos, ou dropamos a coluna inteira.
2- Feature engineering: Nosso foco nessa etapa é valorizar as informações dos dados. Em cada coluna adotamos diferentes abordagens com o intuito de valorizar as informações contidas nos dados.

1- Etapa de pré-processamento:

Dividimos o pré processamento dos dados entre os 4 integrantes do grupo. Cada aluno produziu seu notebook, com os códigos para a manipulação, e no final juntamos todo os códigos para a produção do script final, assim permitindo que o pré processamento seja replicável.

X pode ser: [1 , 2, 3, 4, 5]

notebooks: src / notebooks / preprocessing / 0X_preprocessing.ipynb
script final: src / scripts / data_preprocessing.py

01_preprocessing.py

Nome da coluna: PESSOA_PIPEDRIVE_birthdate
Manipulação: Conversão para datetime.
Descrição: A coluna 'PESSOA_PIPEDRIVE_birthdate' foi convertida para o formato datetime usando pd.to_datetime(), a fim de facilitar cálculos relacionados a datas.

Nome da coluna: PESSOA_PIPEDRIVE_age
Manipulação: Cálculo da idade.
Descrição: A idade foi calculada a partir da data de nascimento, considerando o ano atual e verificando se o aniversário já ocorreu no ano corrente.

Nome da coluna: PESSOA_PIPEDRIVE_age (Continuação)
Manipulação: Imputação de valores faltantes e arredondamento.
Descrição: Utilizando SimpleImputer com estratégia de 'mean', os valores faltantes foram substituídos pela média. Após isso, os valores foram arredondados e convertidos para inteiros.

Nome da coluna: PESSOA_PIPEDRIVE_birthdate
Manipulação: Remoção da coluna.
Descrição: A coluna original de data de nascimento foi removida após o cálculo da idade.

Nome da coluna: PESSOA_PIPEDRIVE_id_gender
Manipulação: Exclusão de linhas e imputação de valores faltantes.
Descrição: Linhas com valores específicos em 'PESSOA_PIPEDRIVE_id_gender' foram removidas. Valores faltantes foram substituídos pelo código 64.

Nome da coluna: PESSOA_PIPEDRIVE_id_marrital_status
Manipulação: Imputação de valores faltantes.
Descrição: Valores faltantes em 'PESSOA_PIPEDRIVE_id_marrital_status' foram preenchidos com o código 80.
Nome da coluna: PESSOA_PIPEDRIVE_state
Manipulação: Imputação de valores faltantes.
Descrição: Valores faltantes foram substituídos pela moda (valor mais frequente) da coluna.

Nome da coluna: PESSOA_PIPEDRIVE_city
Manipulação: Imputação de valores faltantes.
Descrição: Valores faltantes foram substituídos pela moda da coluna.

Nome da coluna: PESSOA_PIPEDRIVE_id_health_plan

Manipulação: Imputação de valores faltantes e remoção da coluna.
Descrição: Valores faltantes foram substituídos pela moda. Após a criação de uma nova coluna baseada nessa, a coluna original foi removida.

Nome da coluna: PESSOA_PIPEDRIVE_has_public_health_plan
Manipulação: Criação de coluna baseada em condição.
Descrição: Nova coluna criada para indicar se o plano de saúde é público (valor 412) com valores binários (1 ou 0).

Nome da coluna: PESSOA_PIPEDRIVE_tem_data_de_termino_de_contrato
Manipulação: Criação de coluna baseada na existência de data.
Descrição: Nova coluna criada para indicar a existência de uma data de término de contrato, usando valores binários (1 ou 0).

Nome da coluna: PESSOA_PIPEDRIVE_id_continuity_pf
Manipulação: Remoção da coluna.
Descrição: A coluna 'PESSOA_PIPEDRIVE_id_continuity_pf' foi removida do conjunto de dados.

Nome da coluna: PESSOA_PIPEDRIVE_Canal de Preferência
Manipulação: Imputação de valores faltantes.
Descrição: Valores faltantes na coluna foram preenchidos com 0.

Nome da coluna: PESSOA_PIPEDRIVE_Tem_Canal_de_Preferência
Manipulação: Criação de coluna baseada em condição.
Descrição: Nova coluna criada para indicar a existência de um canal de preferência, utilizando valores binários (1 ou 0).

Nome da coluna: PESSOA_PIPEDRIVE_has_notes
Manipulação: Criação de coluna baseada em condição.
Descrição: Nova coluna criada para indicar a existência de anotações, com valores binários (1 ou 0) baseados na contagem de anotações.

02_preprocessing.ipynb

Nome da coluna: ATENDIMENTOS_AGENDA_Faltas Psicoterapia
Manipulação: Coluna removida.
Descrição: A coluna, que contava faltas em sessões de psicoterapia, foi removida devido à baixa quantidade de valores não-nulos (menos de 40%).

Nome da coluna: TWILIO_Ligações Inbound
Manipulação: Coluna removida.

Descrição: A coluna, que registrava o número de ligações recebidas, foi removida por ter menos de 40% de valores não-nulos.

Nome da coluna: TWILIO_Data Última Ligações Inbound
Manipulação: Coluna removida.
Descrição: A coluna, indicando a data da última ligação recebida, foi removida devido à baixa quantidade de dados válidos.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Faturas
Manipulação: Coluna removida.
Descrição: A coluna, contando faturas emitidas para associados PF, foi removida por ter uma quantidade insuficiente de dados válidos.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Tentativas de Cobrança
Manipulação: Coluna removida.
Descrição: A coluna, registrando tentativas de cobrança para associados PF, foi removida por falta de dados suficientes.

Nome da coluna: COBRANÇA_VINDI_Método de Pagamento
Manipulação: Coluna removida.
Descrição: A coluna, que detalhava os métodos de pagamento de associados PF, foi removida devido à baixa quantidade de dados válidos.

Nome da coluna: COBRANÇA_VINDI_Valor Médio da Mensalidade
Manipulação: Coluna removida.
Descrição: A coluna, indicando a média do valor mensal cobrado de associados PF, foi removida por ter poucos dados válidos.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Faturas Pagas após Vencimento
Manipulação: Coluna removida.
Descrição: A coluna, que contava faturas pagas após o vencimento por associados PF, foi removida devido à insuficiência de dados.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Faturas Inadimpletes
Manipulação: Coluna removida.
Descrição: A coluna, que registrava a quantidade de faturas inadimplentes de associados PF, foi removida por falta de dados suficientes.

Nome da coluna: COBRANÇA_VINDI_Valor Total Inadimplência
Manipulação: Coluna removida.
Descrição: A coluna, somando o valor das faturas vencidas de associados PF, foi removida por ter uma quantidade insuficiente de dados válidos.

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Psicoterapia
Manipulação: Preenchimento de valores nulos.
Descrição: Os valores Nan foram interpretados como "não realizou sessões de psicoterapia", portanto, foram preenchidos com 0.

Nome da coluna: ATENDIMENTOS_AGENDA_Datas Psicoterapia
Manipulação: Criação de nova coluna.
Descrição: Foi criada a coluna "ATENDIMENTOS_AGENDA_Datas Psicoterapia Recente" que informa se a pessoa realizou uma sessão de psicoterapia em um período mais recente que a mediana da coluna.

Nome da coluna: WHOQOL_Qde Respostas WHOQOL
Manipulação: Coluna removida.
Descrição: A coluna foi removida por não ter relevância para o modelo.

Nome da coluna: WHOQOL_Físico
Manipulação: Criação de nova coluna e tratamento de valores.
Descrição: A última nota foi selecionada quando havia múltiplas notas. Os valores nulos foram preenchidos com a mediana em uma nova coluna "WHOQOL_Físico_New".

Nome da coluna: WHOQOL_Psicológico
Manipulação: Criação de nova coluna e tratamento de valores.
Descrição: A última nota foi selecionada em casos de múltiplas notas. Valores nulos foram preenchidos com a mediana na nova coluna "WHOQOL_Psicológico_New".

Nome da coluna: WHOQOL_Social
Manipulação: Criação de nova coluna e tratamento de valores.
Descrição: A última nota foi usada em casos de múltiplas notas. Valores nulos foram preenchidos com a mediana na nova coluna "WHOQOL_Social_New".

Nome da coluna: WHOQOL_Ambiental
Manipulação: Criação de nova coluna e tratamento de valores.
Descrição: Em casos de múltiplas notas, a última foi selecionada. Valores nulos foram preenchidos com a mediana na nova coluna "WHOQOL_Ambiental_New".

Nome da coluna: COMUNICARE_Problemas Abertos
Manipulação: Criação de nova coluna booleana.
Descrição: Foi criada uma coluna booleana onde o valor é 1 se houve queixa e 0 caso contrário, enfatizando a presença de queixas em vez do conteúdo específico.

Nome da coluna: TWILIO_Mensagens Inbound
Manipulação: Preenchimento de valores nulos.

Descrição: Valores Nan foram interpretados como 0 mensagens enviadas pela pessoa, sendo preenchidos com o valor 0.

Nome da coluna: TWILIO_Data Última Mensagens Inbound
Manipulação: Criação de nova coluna.
Descrição: Foi criada a coluna "TWILIO_Data Última Mensagens Inbound Recente", que informa se a pessoa enviou uma mensagem recentemente, com base na mediana da coluna.

03_preprocessing.py

Nome da coluna: FUNIL_ASSINATURA_PIPEDRIVE_lost_time
Manipulação: Preenchimento e substituição de valores.
Descrição: Se a data de fim da assinatura ('lost_time') é nula e existe um contrato finalizado, insere-se a data de fim do contrato. Se 'lost_time' é nulo e o contrato está em andamento, substitui por "em aberto". Se há mais de uma data, usa-se a primeira e normaliza-se no formato YYYY-MM-DD.

Nome da coluna: FUNIL_ASSINATURA_PIPEDRIVE_lost_reason
Manipulação: Classificação e agrupamento.
Descrição: Mantém o primeiro motivo de cancelamento da assinatura listado. Agrupa motivos menos frequentes em "Outro" e classifica valores nulos como "Assinatura ativa".

Nome da coluna: FUNIL_ONBOARDING_PIPEDRIVE_add_time
Manipulação: Preenchimento de valores nulos.
Descrição: Valores nulos são preenchidos como "não iniciado", indicando que o processo de onboarding não começou.

Nome da coluna: FUNIL_ONBOARDING_PIPEDRIVE_status
Manipulação: Substituição de valores faltantes.
Descrição: Substitui valores faltantes por "não iniciado", aplicável quando 'add_time' é nulo.

Nome da coluna: FUNIL_ONBOARDING_PIPEDRIVE_activities_count
Manipulação: Tratamento de valores nulos.
Descrição: Valores nulos são tratados como 0, indicando que nenhuma atividade de onboarding foi concluída.

Nome da coluna: FUNIL_ONBOARDING_PIPEDRIVE_lost_reason
Manipulação: Preenchimento e categorização.

Descrição: Preenche valores nulos como "processo em aberto", "processo concluído" ou "processo não iniciado", dependendo de outras colunas. Agrupa valores menos frequentes em "outros".

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Atendimento Médico
Manipulação: Tratamento de nulos.
Descrição: Trata valores nulos como 0, representando a ausência de atendimentos médicos.

Nome da coluna: ATENDIMENTOS_AGENDA_Faltas Atendimento Médico
Manipulação: Preenchimento de nulos.
Descrição: Valores nulos são tratados como 0, indicando que não houve faltas.

Nome da coluna: ATENDIMENTOS_AGENDA_Datas Atendimento Médico
Manipulação: Preenchimento de nulos.
Descrição: Trata valores nulos como “nunca ocorreu”, indicando a ausência de atendimentos.

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Atendimentos Acolhimento
Manipulação: Preenchimento de nulos.
Descrição: Valores nulos são tratados como 0, indicando a ausência de atendimentos de acolhimento.

Nome da coluna: ATENDIMENTOS_AGENDA_Faltas Acolhimento
Manipulação: Preenchimento de nulos.
Descrição: Valores nulos são tratados como 0, significando a ausência de faltas.

Nome da coluna: ATENDIMENTOS_AGENDA_Datas Acolhimento
Manipulação: Preenchimento de nulos.
Descrição: Trata valores nulos como “nunca ocorreu”, indicando a ausência de atendimentos de acolhimento.

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Psicoterapia
Manipulação: Preenchimento de nulos.
Descrição: Valores nulos são tratados como 0, representando a ausência de sessões de psicoterapia.

Nome da coluna: ATENDIMENTOS_AGENDA_Faltas Psicoterapia
Manipulação: Preenchimento de nulos.
Descrição: Valores nulos são tratados como 0, indicando a ausência de faltas em sessões de psicoterapia.

Nome da coluna: ATENDIMENTOS_AGENDA_Datas Psicoterapia
Manipulação: Preenchimento de nulos.

Descrição: Trata valores nulos como “nunca ocorreu”, indicando a ausência de sessões de psicoterapia.

Nome da coluna: stay_time (feature engineering)
Manipulação: Criação de nova coluna.
Descrição: Calcula o tempo total do usuário na plataforma usando 'lost_time' - 'start_of_service' ou 'lost_time' - 'contract_start_date', dependendo da disponibilidade dos dados.

Nome da coluna: last_stage_concluded (feature engineering)
Manipulação: Criação de nova coluna.
Descrição: Usa informações de diversas colunas ('stay_in_pipeline_stages_welcome', 'first_meeting', 'whoqol') para determinar o último processo concluído. As colunas usadas são removidas após a criação desta.

Nome da coluna: process_time (feature engineering)
Manipulação: Criação de nova coluna.
Descrição: Combina 'won_time' e 'lost_time' em uma coluna. Utiliza 'lost_time' se disponível; caso contrário, 'won_time'. Se 'add_time' for nulo, o valor é "não iniciado"; senão, "em aberto". Colunas originais são removidas após a criação desta.

04_preprocessing.py

Nome da coluna: ATENDIMENTOS_AGENDA_Faltas Psicoterapia
Manipulação: Remoção de Coluna.
Descrição: A coluna que contabilizava as faltas em sessões de psicoterapia foi removida.

Nome da coluna: TWILIO_Ligações Inbound
Manipulação: Remoção de Coluna.
Descrição: A coluna que registrava o número de ligações recebidas foi removida.

Nome da coluna: TWILIO_Data Última Ligações Inbound
Manipulação: Remoção de Coluna.
Descrição: A coluna que indicava a data da última ligação recebida foi removida.

Nome da coluna: TWILIO_Data Última Mensagens Outbound
Manipulação: Remoção de Coluna.
Descrição: A coluna que registrava o tempo desde a última mensagem enviada pela equipe de saúde foi removida.

Nome da coluna: TWILIO_Data Última Mensagens Outbound Tempo passado
Manipulação: Remoção de Coluna.
Descrição: A coluna que indicava a data da última mensagem enviada foi removida.

Nome da coluna: TWILIO_Data Última Ligações Outbound
Manipulação: Remoção de Coluna.
Descrição: A coluna que contava as ligações feitas pela equipe de saúde foi removida.

Nome da coluna: TWILIO_Data Última Ligações Outbound Tempo passado
Manipulação: Remoção de Coluna.Descrição: A coluna que indicava a data da última ligação feita pela equipe foi removida.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Faturas
Manipulação: Remoção de Coluna.
Descrição: A coluna que registrava o total de faturas emitidas foi removida.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Tentativas de Cobrança
Manipulação: Remoção de Coluna.
Descrição: A coluna que contabilizava o total de tentativas de cobrança foi removida.

Nome da coluna: COBRANÇA_VINDI_Método de Pagamento
Manipulação: Remoção de Coluna.
Descrição: A coluna que detalhava os métodos de pagamento utilizados foi removida.

Nome da coluna: COBRANÇA_VINDI_Valor Médio da Mensalidade
Manipulação: Remoção de Coluna.
Descrição: A coluna que indicava a média do valor mensal cobrado foi removida.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Faturas Pagas após Vencimento
Manipulação: Remoção de Coluna.
Descrição: A coluna que registrava o total de faturas pagas após o vencimento foi removida.

Nome da coluna: COBRANÇA_VINDI_Qde Total de Faturas Inadimplentes
Manipulação: Remoção de Coluna.
Descrição: A coluna que contava o total de faturas inadimplentes foi removida.

Nome da coluna: COBRANÇA_VINDI_Valor Total Inadimplência
Manipulação: Remoção de Coluna.

Descrição: A coluna que somava o valor total das inadimplências foi removida.

Nome da coluna: COBRANÇA_VINDI_Qde Perfis de Pagamento Inativos
Manipulação: Remoção de Coluna.
Descrição: A coluna que contabilizava o total de perfis de pagamento inativos foi removida.

Nome da coluna: TWILIO_Data Última Mensagens Outbound Recente
Manipulação: Criação de Nova Coluna.
Descrição: Criada a partir das colunas removidas sobre mensagens outbound, essa coluna indica

Etapa de Feature engineering:

Dividimos a feature engineering dos dados entre os integrantes do grupo. Assim produzindo notebooks

X pode ser: [1 , 2, 3, 4]

notebooks: src / notebooks / feature enginering / 0X_preprocessing.ipynb

script final: src / scripts / data_feature_enginering.py

01_feature_engineering.ipynb

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Atendimentos Acolhimento
Manipulação: Criação de nova coluna.
Descrição: Baseado na contagem dos atendimentos realizados com a equipe de acolhimento da Ana Health, e em conjunto com a coluna "ATENDIMENTOS_AGENDA_Datas Acolhimento", foi criada uma nova coluna indicando a frequência de atendimentos por mês de cada cliente.

Nome da coluna: ATENDIMENTOS_AGENDA_Faltas Acolhimento
Manipulação: Criação de nova coluna.
Descrição: Em combinação com a coluna "ATENDIMENTOS_AGENDA_Qde Atendimentos Acolhimento", foi criada uma nova coluna para indicar a taxa de falta

de cada cliente, ajudando a identificar o nível de comprometimento do cliente com os agendamentos.

Nome da coluna: TWILIO_Mensagens Inbound e TWILIO_Mensagens Outbound
Manipulação: Cálculo de razão.
Descrição: A partir destas duas colunas, foi calculada a razão entre as mensagens recebidas e enviadas, refletindo o engajamento e a participação ativa do cliente na plataforma.

Nome da coluna: PESSOA_PIPEDRIVE_age
Manipulação: Categorização por faixa etária.
Descrição: A coluna foi dividida em categorias representando faixas etárias (criança, jovem, adulto, idoso), para melhor segmentação e compreensão dos dados demográficos.

Nome da coluna: TWILIO_Ligações Outbound
Manipulação: Criação de nova coluna.
Descrição: Foi criada uma nova coluna indicando se o cliente recebeu uma quantidade significativa de ligações da plataforma, visando identificar clientes potencialmente mais propensos a abandonar a plataforma.

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Psicoterapia
Manipulação: Segmentação de clientes.
Descrição: Os clientes foram segmentados com base na quantidade de agendamentos de psicoterapia, criando faixas como "nenhum agendamento", "poucos agendamentos", "muitos agendamentos", etc., para personalização da abordagem.

Nome da coluna: ATENDIMENTOS_AGENDA_Qde Prescrições e ATENDIMENTOS_AGENDA_Datas Prescrição
Manipulação: Preenchimento de valores e remoção de coluna.
Descrição: A coluna de quantidade de prescrições foi preenchida com 0, e a coluna de datas das prescrições foi removida.

Nome da coluna: PESSOA_PIPEDRIVE_id_person
Manipulação: Remoção da coluna.
Descrição: A coluna identificadora única da pessoa no Pipedrive foi removida.

Nome da coluna: PESSOA_PIPEDRIVE_id_gender
Manipulação: Criação de coluna binária e exclusão de linhas.
Descrição: Foram selecionados apenas os gêneros mais significativos (masculino e feminino), e criada uma nova coluna binária representando o gênero do cliente (1 para masculino, 0 para feminino).

Nome da coluna: PESSOA_PIPEDRIVE_id_marrital_status
Manipulação: One-hot encoding.
Descrição: Foi realizado o one-hot encoding para transformar o identificador do estado civil da pessoa em múltiplas colunas binárias.

Nome da coluna: PESSOA_PIPEDRIVE_state
Manipulação: One-hot encoding.
Descrição: O estado de residência da pessoa foi transformado em múltiplas colunas binárias utilizando one-hot encoding.

Nome da coluna: PESSOA_PIPEDRIVE_city
Manipulação: Codificação de frequência.
Descrição: A cidade de residência foi codificada com base na frequência de ocorrência de cada cidade no conjunto de dados.

02_feature_engineering.ipynb

Nome da coluna: stay_time
Manipulação: Criação de nova coluna.
Descrição: Calcula o tempo total do usuário na plataforma. Utiliza a diferença entre 'lost_time' e 'start_of_service' ou 'lost_time' e 'contract_start_date', dependendo da disponibilidade de 'start_of_service'.

Nome da coluna: last_stage_concluded
Manipulação: Criação de nova coluna e remoção das colunas originais.
Descrição: A partir das informações das colunas 'stay_in_pipeline_stages_welcome', 'stay_in_pipeline_stages_first_meeting' e 'stay_in_pipeline_stages_whoqol', foi criada uma nova coluna que indica o último processo concluído pelo usuário. As colunas utilizadas para a criação desta foram removidas posteriormente.

Nome da coluna: process_time
Manipulação: Combinação de colunas e remoção das colunas originais.
Descrição: Combina as informações de 'won_time' e 'lost_time' em uma única coluna. Se 'lost_time' estiver disponível, é utilizado; se 'won_time' estiver disponível, é utilizado. Se 'add_time' for nulo, o valor atribuído é "não iniciado"; caso contrário, "em aberto". As colunas utilizadas para a criação desta foram removidas.

Nome da coluna: FUNIL_ASSINATURA_PIPEDRIVE_status
Manipulação: Aplicação de one-hot encoding.
Descrição: One hot encoding foi aplicado, gerando novas colunas “assinatura_status_won” e “assinatura_status_lost”.

Nome da coluna: FUNIL_ONBOARDING_PIPEDRIVE_status
Manipulação: Aplicação de one-hot encoding.
Descrição: One hot encoding foi aplicado nos dados da coluna, resultando em novas colunas: assinatura_status_Não iniciado, assinatura_status_lost, assinatura_status_open e assinatura_status_won.

Nome da coluna: FUNIL_ONBOARDING_PIPEDRIVE_lost_reason
Manipulação: Aplicação de one-hot encoding.
Descrição: One hot encoding foi aplicado, criando colunas como onboarding_lost_reason_Outro, onboarding_lost_reason_[Onboarding] Não retornou aos contatos de resgate, entre outras.

Nome da coluna: FUNIL_ASSINATURA_PIPEDRIVE_lost_reason
Manipulação: Aplicação de one-hot encoding.
Descrição: One hot encoding foi aplicado, gerando colunas como assinatura_lost_reason_Outro, assinatura_lost_reason_[Assinatura] Cancelamento por inadimplência, entre outras.

Nome da coluna: FUNIL_ASSINATURA_PIPEDRIVE_start_of_service
Manipulação: Remoção da coluna.
Descrição: A coluna foi removida após ser utilizada para a construção de outras colunas.

Nome da coluna: FUNIL_ASSINATURA_PIPEDRIVE_lost_time
Manipulação: Remoção da coluna.
Descrição: A coluna foi removida após ser utilizada para a construção de outras colunas.

03_feature_engineering.ipynb

Nome da coluna: PESSOA_PIPEDRIVE_contract_start_date
Manipulação: Remoção da coluna.
Descrição: A coluna 'PESSOA_PIPEDRIVE_contract_start_date', que indicava a data de início do contrato, foi removida após sua utilização na construção de outras colunas.

Nome da coluna: PESSOA_PIPEDRIVE_contract_end_date
Manipulação: Remoção da coluna.
Descrição: A coluna 'PESSOA_PIPEDRIVE_contract_end_date', que representava a data de término do contrato, foi removida do conjunto de dados após ter sido utilizada para a criação de outras colunas.

Nome da coluna: PESSOA_PIPEDRIVE_Canal de Preferência
Manipulação: Aplicação de one-hot encoding.
Descrição: One hot encoding foi aplicado nos dados da coluna 'PESSOA_PIPEDRIVE_Canal de Preferência', resultando na criação de novas colunas denominadas canal_de_preferencia_0.0, canal_de_preferencia_238.0, canal_de_preferencia_239.0, canal_de_preferencia_360.0. Cada nova coluna representa uma categoria única dentro da coluna original.

04_feature_engineering.ipynb

Nome da coluna: PESSOA_PIPEDRIVE_has_notes
Manipulação: Verificação da distribuição.
Descrição: Como 'PESSOA_PIPEDRIVE_has_notes' é uma variável binária (0 ou 1), foi realizada uma análise da distribuição dos valores para entender a proporcionalidade e avaliar a necessidade de técnicas de reamostragem ou ponderação para balancear a influência dessa feature no modelo. Não foi identificada a necessidade de balanceamento.

Nome da coluna: WHOQOL_Físico_New, WHOQOL_Psicológico_New, WHOQOL_Social_New
Manipulação: Aplicação de one-hot encoding.
Descrição: Como são variáveis categóricas, foi aplicada a codificação one-hot para transformá-las em variáveis binárias, facilitando a utilização em modelos preditivos.

Nome da coluna: COMUNICARE_Problemas Abertos Bool
Manipulação: Verificação da distribuição.
Descrição: Similar à variável 'PESSOA_PIPEDRIVE_has_notes', esta também é uma variável binária. Foi realizada uma análise de distribuição para verificar a necessidade de balanceamento, mas não foi identificado desequilíbrio significativo.

Nome da coluna: TWILIO_Data Última Mensagens Inbound Recente, TWILIO_Data Última Mensagens Outbound Recente, TWILIO_Data Última Ligações Outbound Recente
Manipulação: Conversão de datas para formato numérico.
Descrição: As datas foram convertidas em um formato numérico, como o número de dias desde uma data de referência, criando features que indicam o tempo desde a última interação.

Nome da coluna: stay_time
Manipulação: Análise de distribuição.
Descrição: Foi realizada uma análise de distribuição para a feature 'stay_time', mas não foi identificada a necessidade de balanceamento.

Nome da coluna: last_stage_concluded
Manipulação: Aplicação de one-hot encoding.
Descrição: Sendo uma variável categórica que indica diferentes estágios, foi aplicada a codificação one-hot para transformar cada estágio em uma variável binária.

Nome da coluna: process_time
Manipulação: Transformação e tratamento de outliers.
Descrição: Foram aplicadas transformações apropriadas com base na distribuição dos dados. A coluna, sendo uma medida de tempo, teve seus outliers identificados e tratados, principalmente removendo aqueles registros que estavam em aberto.