

Framework de avaliação - Heitor Felix de Oliveira

a. Quantos contratos distintos existem em sua tabela de observação final, após todos os merges e etapas de ETL?

14.756 contratos distintos

b. Você considerou ou aplicou algum filtro de exclusão? Caso afirmativo, descreva-o e informe quantos contratos foram filtrados.

Alguns filtros foram aplicados no portfólio geral e no portfólio tpv. Todos valores negativos foram apagados. Mas nenhum contrato foi excluído, somente alguns registros com valores estranhos.

- Linhas removidas para valores de pagamento negativo: 23
- Linhas removidas para valores de pagamento negativo: 1591
- Linhas removidas para valores de quantidade de transação negativo: 7
- Linhas removidas para valores de transação de transação negativo: 1238

Na tabela com contratos únicos foram desconsiderados linhas de contrato duplicadas com valor de tipo de empresas diferentes, devido ao pouco impacto que teria.

Etapas 1 - Limpeza dos dados - (Python)

- Para a tabela *portfolio_geral* foram selecionadas somente as colunas que variam e chaves primárias (contrato_id (PK), dt_ref_portfolio (PK), nr_documento, status_contrato, perc_retencao, vlr_pgto_realizado, vlr_pgto_esperado, vlr_saldo_devedor, vlr_saldo_devedor_esperado, dsp, dspp, flag_transacao, qtd_transacoes, vlr_tpv). A tabela filtrada passou a ser chamada de *df_geral*;
- As colunas que eram constantes, foram separadas em outra tabela chamada *df_contrato*. Os duplicados foram removidos e esse dataframe contém todas as informações do contrato, chaves primárias e estrangeiras (contrato_id (PK), nr_documento (FK), safra, dt_contrato, dt_desembolso, dt_vencimento, dt_wo, prazo, vlr_desembolsado, vlr_tarifa, juros_mes, juros_diario);
- Foram removidas todas as linhas que tinham *status_contrato* = 'Settled' exceto a primeira, pois uma vez que o empréstimo foi quitado, os próximos registros não interessam. Cerca de 25% das linhas foram removidas aqui, sem prejuízo de informação;

- Valores negativos de quantidade e valor foram removidos (considerando o total, pouca informação foi removida). Tanto nas tabelas *df_geral* e *portfolio_tpv*.

Objetivo: A limpeza foi feita principalmente para reduzir o número de linhas e melhorar a performance do script e do dashboard e também para excluir valores esquisitos que poderiam afetar a análise.

Etapa 2 - Join de tabelas - (Python)

- Join da tabela *df_geral* com o *portfolio_comunicados* pelo *contrato_id* e *dt_ref_portfolio*, para extrair os comunicados (se houve) em cada registro histórico do contrato (tipo left join). Onde não houve comunicação, ficou com valores de ação nulos e alguns registros foram duplicados pois houveram mais de uma ação no mesmo dia para o mesmo contrato, foi criado portanto uma nova tabela chamada *df_geral_comunicado* que recebeu esse join;
- Join da tabela *df_contrato* com o *portfolio_clientes* pelo *nr_documento*, para extrair as informações do cliente de cada contrato (tipo inner join). Essa foi armazenada no próprio *df_contrato*;
- Join da tabela *df_geral* com *portfolio_tpv* pelo *nr_documento* e *dt_transacao* com *dt_ref_portfolio*, para extrair as informações de transação pela maquininha em cada registro histórico do contrato (tipo left join), valores nulos foram substituídos por 0, pois significa que houve 0 transação.

Objetivo: Conseguir cruzar informações e correlacionar variáveis que estavam separadas.

Etapa 3 - Criação de novas tabelas - Python

- As colunas que eram constantes, foram separadas em outro dataframe chamado *df_contrato*. Os duplicados foram removidos e esse dataframe contém todas as informações do contrato. Algumas linhas tinham valores de *contrato_id* duplicado com informações de tipo de empresa diferente, mas como eram poucos casos, os duplicados foram removidos;
- Uma tabela chamada *df_mensagem* foi criada apenas com as linhas do *df_geral_comunicado* que continha uma comunicação e os 5 registros seguintes. O objetivo desse dataframe é observar o efeito da comunicação nos 5 dias seguintes, através do *dsp* e *dspp*. Dessa vez o enfoque é descobrir a curva ideal de vezes que se deve acionar o cliente.

Objetivo: A tabela de contratos teve o objetivo de sumarizar todas as informações referentes a cada contrato, além de resumir as informações que se repetiam no portfólio geral. Já a tabela de mensagem teve o objetivo de analisar os acionamentos e a eficiência dos comunicados.

Etapa 4 - Criação de novas features - Python

4.1 df_geral

Variável	Descrição/Cálculos
valor_pago_tpv	Valor (tpv) * percentual de retenção
valor_total_pago	vlr_pgto_realizado + valor_pago_tpv

4.2. df_mensagem

Variável	Descrição/Cálculos
eficiencia	Se o contratante efetua pagamento após mensagem de régua dsp a eficiência é 1, se o contratante efetua pagamento principal após mensagem de régua dspp a eficiência também é 1. Se o contratante lê a ação e continua sem pagar a eficiência é 0

4.3. df_contrato

Variável	Descrição/Cálculos
ultima_data	Última data registrada pelo contrato
status_contrato	Qual o status do contrato no último dia registrado
dsp	média aritmética de dsp, extraída do df_geral
dspp	média aritmética de dspp, extraída do df_geral
valor_pago_tpv	soma do valor pago tpv, extraída do df_geral
vlr_pgto_realizado	soma do valor pago realizado, extraído do df_geral
valor_total_pago	Soma do valor pago tpv + Soma do valor pago realizado, agrupado por contrato
tempo_esperado	Tempo em dias entre data do desembolso e data do vencimento do contrato
dias_de_contrato	Tempo em dias entre primeira data e a última de contrato
n_acoes_total	Número total de ações enviada registradas
n_acoes_entregues	Número total de ações , desconsiderando as que não foram entregues
saldo_devedor_acumulado	$vlr_desembolsado * (1 + juros_diario)^{dias}$
debito	Saldo devedor acumulado - valor total pago
mensagens_lidas	Nº de mensagens com status = "Lido"
eficiencia	Soma de eficiência / nº de mensagens lidas, agrupado por contrato
dias_de_atraso	Dias de contrato ativo após a data de vencimento
tempo_restante	Diferença entre dt_wo e ultima_data em dias
ritmo_esperado	Quanto o contratante teria que pagar por dia para concluir a dívida com o tempo que falta. Débito / tempo restante
pgto_médio	média de valor total pago por dia, agrupado por contrato

Objetivo: Para essa semifinal percebi que algumas features precisavam ser criadas e outras faltavam ser construídas.

Etapa 5 - Clusterização

- Cada contrato recebeu uma categoria de acordo com seu ritmo de pagamento e o engajamento aos acionamentos.
- O cluster dos bons pagadores foi o primeiro a ser criado. Ele foi criado seguindo as seguintes condições: $n_acoes_total < 10$ e média de dias sem pagar < 10 .
- Fiz a seleção de features para rodar o modelo de Machine Learning (ML) para encontrar os dois outros clusters. As features selecionadas foram: *eficiência média*, *nº de mensagens enviadas*, *pagamento médio realizado e esperado*.
- O modelo de ML escolhido foi o KMeans e obtive um silhouette score de 0.47. Os outros dois clusters encontrados foram os *engajados* e os *desengajados*.
- As tabelas com novas features foram exportadas para .csv e foram analisadas em outro script.

Objetivo: As características dos contratos variam muito quanto à resposta aos acionamentos. Dividir melhor os contratos em termos de engajamento e ritmo de pagamento ajuda a separar a análise dos grupos mais engajados e desengajados.

Etapa 6 - Análise de dados 1.0 - Python

Todas as análises estão detalhadas no notebooks, mas estes são os principais insights:

- O valor que ainda está para ser pago é: R\$81,450,085.45;
- O saldo até agora foi negativo com o valor de : R\$-11,707,166.48;
- 27.85% dos contratos estão atrasados;
- Taxa de leitura das ações: 35.09%;

- Taxa de eficiência das ações lidas: 42.16% ;
- No script é possível observar o impacto que teria caso mais mensagens fossem entregues;
- A campanha de Pré Negativação é mais eficiente que a campanha de Boleto Quitado, mesmo sendo uma campanha mais tardia.

Etapa 7 - Análise de dados 2.0 - Python

Nesta etapa a análise foi feita de uma forma geral, analisando o engajamento das mensagens de acordo em diferentes óticas.

4.1. Taxa de mensagens entregues

- Aqui foi analisado se existe alguma diferenciação para entender a taxa de mensagens entregues.

4.2. Taxa de leitura

- Aqui foi analisado se existe alguma diferenciação para entender a taxa de leitura.

4.3. Eficiência média

- A eficiência média das campanhas foram analisadas ao longo dos dias sem pagar.

4.4. Clusters

- Os clusters encontrados foram separados em termos de eficiência média de acionamento da seguinte forma:
 - Engajados \rightarrow eficiencia ≥ 0.4
 - Desengajados \rightarrow eficiencia < 0.4
- Elaboração da nova régua seguindo a *curva de acionamentos ideal* pedido pelo desafio. A régua original é comparada graficamente à régua proposta com suas estimativas de eficiência.

Principais insights:

- Acionamentos por e-mail e Whatsapp performam da mesma maneira em termos de taxa de mensagens entregues, taxa de leitura e eficiência média.
- A taxa de leitura varia muito pouco ao longo do tempo.

- A eficiência dos acionamentos diminui ao longo do tempo, é o atributo mais relevante para o problema, pois é a que mais varia.

Etapa 8 - Criação do Dashboard Power BI

8.1. Curva de engajamento

- **Filtros:** Filtros de ano, de status do contrato e de cluster.
- **Cartões:** Informações gerais de número de contratos ativos (aqueles onde o último registro é do ano analisado), eficiência média e nº médio de mensagens entregues.
- **Gráfico de linha - Curva de engajamento, réguas dsp e dspp:** Esse gráfico mostra como o engajamento varia ao longo dos dias (sob as réguas dsp e dspp). Cada área do gráfico é uma zona de engajamento e os pontos são onde existem acionamentos. A cor da linha indica se o acionamento foi de campanha de pagamento ou de pagamento principal.
- **Tabela** - Essa tabela mostra valores o número de mensagens enviadas, entregues, lidas e eficientes de cada campanha, e por fim mostra a eficiência média da campanha.
- **Gráfico de linha** - Esse gráfico mostra como a eficiência média variou ao longo do tempo. É possível analisar se houve evolução, retrocesso ou constância ao longo do tempo.

Essa página tem como objetivo analisar como o engajamento varia de acordo com outras variáveis.

Principais insights:

- **Régua dsp:** O engajamento diminui ao longo do tempo. Ele tem decrescimento acentuado nos primeiros 15 dias, onde chamei de *zona segura*. Depois ele diminui o ritmo de decrescimento menos acentuado, mas entre 15 e 60 dias está a área em que chamei de *zona preocupante*, onde o engajamento é muito baixo e já é difícil fazer o contratante voltar a pagar. E após os 60 dias sem pagar, é a zona perdida, onde é extremamente difícil a volta do ritmo saudável de pagamento

- Régua dspp: O engajamento segue o mesmo padrão da curva dos acionamentos de régua dsp, portanto é bem menos eficiente por só começar na *zona preocupante*, onde o engajamento já começa baixo e decai mais ainda ao longo do tempo. O engajamento ainda é menor do que os acionamentos que tem régua dsp, provavelmente por ser cobrado valores maiores nestas campanhas.
- O comportamento do engajamento segue a curva do gráfico para todos os clusters, o que muda é a proporção, onde os valores de eficiência média são muito maiores entre os engajados e bons pagadores.
- Engajamento ao longo do tempo:
 - Em 2020 houve uma grande queda no engajamento em março, provavelmente por causa da pandemia, quando muita gente começou a se tornar inadimplente e fechar os empreendimentos. O engajamento voltou a crescer depois de abril e começou a decrescer para o ano seguinte.
 - Em 2021 houve uma queda constante a partir de abril até o final do ano. Esse comportamento pode ser explicado por conta dos prejuízos ainda da pandemia.
 - Em 2022 o engajamento tem sido constante perto de 0.40. O que pode ser melhorado, já que a pandemia aliviou com a vacinação e os empreendimentos já puderam retomar atividades.

8.2. Monitoramento dos contratantes

- **Filtros:** Filtro de dsp atual (quantos dias o contratante está sem pagar no momento), cluster, situação do contrato, *zona de desengajamento* (5 a 30 dsp) , *zona preocupante* (15 a 60 dsp) , *zona perdida* (+ de 60 dsp). Somente contratos ativos aparecem nessa página.
- **Cartões:** Informações gerais de número de contratos ativos (aqueles onde o último registro é do ano analisado) e soma do débito atual.
- **Tabela:** Informações dos contratos mediante filtros. Estão apresentados o id, dias sem pagar (até última data registrada), ritmo de pagamento diário

esperado e realizado, dias até o write-off do contrato, cluster e eficiência média das mensagens ao contratante. As informações de ritmo de pagamento diário e eficiência são resumidas em média, e o débito é resumido em soma, na última linha.

Esse gráfico tem como objetivo o monitoramento dos inadimplentes, a tabela é ordenada por débito em ordem decrescente. Assim é possível localizar os clientes que estão em diferentes zonas de engajamento e é possível personalizar o contato para cada um deles. Várias listas podem ser extraídas baseado nos filtros selecionados, priorizando variáveis de interesse do consultor. É a principal ferramenta para atender a nova régua de acionamentos que será proposta a seguir.

Principais insights:

- A maior parte do débito (230 milhões) e dos contratos (2,8 mil) está concentrado na *zona preocupante* onde os contratantes estão há mais de 60 dias sem pagar.
- A maior parte do débito (136 milhões) e dos contratos (1,8 mil) são contratos vencidos.

8.3. Perfil dos contratantes

- **Filtros:** Filtros de status do contrato, situação do contrato e cluster.
- **Gráfico de barra:** Compara o número de contratos por tipo de contrato.
- **Gráfico de barra:** Compara o número de contratos por segmento do contratante.
- **Tabela** - Essa tabela mostra o número de contratos por segmento, subsegmento, ritmo de pagamento esperado, realizado e débito. O ritmo de pagamento está pintado de azul conforme o ritmo realizado estiver próximo do esperado. Quanto mais próximo está do esperado, mais azul.
- **Mapa:** Mapa com tamanho de bolhas proporcional à razão (*pagamento diário realizado / pagamento diário esperado*).

O objetivo desta página do dashboard é secundário, é analisar o perfil dos contratantes de diferentes situações, sendo cluster, status e situação do contrato.

Principal insights:

- A proporção do número de contratantes de diferentes segmentos variam por cluster.
- Contratantes de diferentes estados pagam ritmos variados.

8.4. Funil das mensagens e cenários

- **Parâmetros variáveis:** Taxa mensagens entregues, taxa de leitura das mensagens, e taxa de eficiência. Cada parâmetro pode ser alterado pelo usuário para simular cenários onde essas taxas aumentam.
- **Gráfico de barra:** Este gráfico representa o funil dos acionamentos, desde que são enviados até serem lidos e funcionarem. Somente a barra de mensagens enviadas é constante, as outras barras são calculadas através da taxa multiplicada pelo valor da barra anterior.
- **Filtro:** Filtro de cluster, de tipo de ação e de régua.

O objetivo desta página é simular cenários onde podemos ver como seria o comportamento do funil se alterássemos as taxas de envio, leitura e eficiência. O fim do funil são as mensagens que fizeram o contratante voltar a pagar. É possível analisar cada ação, como seria se eles fossem mais eficientes.

O dashboard está disponível neste [link](#):

Conclusão: qual é a curva ideal de vezes que devemos acionar um cliente?

Sobre a eficiência do sistema, foi detectado um problema de alta taxa de comunicados não entregues. A cada 100 mensagens enviadas, cerca de 53 são entregues, das entregues, 17 são lidas e das lidas, 8 convertem em pagamento. Ou seja, muita mensagem é perdida no funil de acionamentos.

Analisando a curva de desengajamento ao longo dos dias sem pagamentos, percebe-se que o engajamento máximo está nos primeiros 15 dias, sofrendo forte queda nesse período. Os acionamentos de dspp deveriam começar mais cedo, para aproveitar esse período de maior engajamento. Portanto, propus uma nova disposição de acionamentos onde as primeiras campanhas de ambas réguas se concentrasse mais no início da curva de engajamento, onde ele é maximizado.

Além disso, propus uma nova campanha de parcelamento entre as campanhas de Pré-negativação e Negativação, na marca de 45 dias sem pagamento, com objetivo de tentar recuperar o contratante antes da negativação com o parcelamento que é a campanha mais eficiente depois da observação.

Para melhorar a eficiência do funil de acionamentos, a melhor estratégia é usar a nova curva de acionamentos, aproveitando o período de maior engajamento e consequentemente a eficiência média das mensagens. Além disso, é importante melhorar o alcance das mensagens através da taxa de mensagens entregues e da taxa de leitura das mensagens, para máxima eficiência dos acionamentos.