**Documento de Resolução do Case**

Este documento faz parte da Resolução do Case da 3ª Etapa Classificatória do Data Challenge 2022 oferecido pela STONE.

O principal intuito aqui é representar de forma sumarizada minha linha de raciocínio e análise durante a resolução do case e garantir uma avaliação mais completa do meu trabalho, além de apresentar o processo de análise que foi feito e de explicar o processo de insights.

1. **Validações iniciais**

Com intuito de validar algumas premissas básicas de junção e preparação dos dados, abaixo seguem algumas perguntas e respostas:

1. Quantos contratos distintos existem na tabela de observação final, após todos os merges e etapas de ETL?

Resposta: Do total de 14.756 contratos distintos, apenas 11.897 contratos distintos existem na tabela de observação final (que contém somente os comunicados que foram entregues, lidos ou respondidos).

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

Resposta: Sim. Para a análise da curva de engajamento x quantidade de acionamentos (comunicados), foram excluídos os contratos que não tiverem nenhum comunicado enviado, ou seja, do total de 14.756 contratos apenas 12.202 contratos tiveram algum comunicado. E do total de contratos que receberam algum comunicado (12.202), 305 contratos deveriam receber algum comunicado, mas, por algum motivo, o comunicado não foi entregue. Assim, apenas 11.897 contratos (12.202 – 305) tiveram comunicados entregues, lidos ou respondidos. E, portanto, 2.859 contratos foram excluídos desta análise.

**Fluxo de etapas realizadas**

Com o intuito de registrar cada etapa do processo realizado durante a solução do case, abaixo seguem as etapas realizadas com todos os detalhes necessários para que qualquer pessoa possa entender a minha abordagem na resolução do problema.

* 1. **Etapa 1 – Análise iniciais**

1. Verifiquei que não há dados faltantes em nenhuma tabela.
2. Sobre os tipos de dados de cada variável, verifiquei que todas as variáveis estão de acordo com o dicionário de dados com exceção da variável “dt\_transacao” da tabela “portfolio\_tpv” que deveria ser do tipo data e está como tipo texto (ou inteiro como identificado pelo Python).
3. Verifiquei que a tabela “portfolio\_clientes” possuem dados duplicados na variável “nr\_documento. Por ser uma tabela de dados cadastrais de clientes, ela deve possuir valores únicos na variável “nr\_documento”, que é a chave primária dessa tabela.
4. Verifiquei que a tabela “portfolio\_tpv” não possui dados duplicados.
5. Na tabela “portfolio\_comunicados”, verifiquei que para os registros que possuem os mesmos valores nas variáveis “contrato\_id”, “dt\_ref\_portfolio”, “data\_acao”, e “tipo\_acao” existem campanhas diferentes de comunicação. Ou seja, para o mesmo contrato e no mesmo dia foram enviadas mais de uma comunicação (campanha). Através da análise de um caso, contrato\_id = “000c35a61297edadc2842f6d5b4028e1” (que possui DPS = 30 e DSPP = 30), verifiquei que este cliente recebeu pelo menos duas ações de comunicação por causa dos valores de DSP e DSPP e a régua de comunicação. Logo, estou assumindo que isso acontece por causa dos momentos diferentes da jornada do contrato do cliente (como realçado na descrição do case).
6. Verifiquei que a tabela “portfolio\_geral” não possui dados duplicados.
   1. **Etapa 2 – Preparação e Limpeza dos dados**
7. Corrigi o tipo de dado da variável “dt\_transacao” da tabela “portfolio\_tpv” para o mesmo formato das outras datas (“aaaa-mm-dd”).
8. Removi valores duplicados na tabela “portfolio\_clientes” mantendo a última ocorrência. Estou assumindo que são atualizações de dados, onde houve uma adição de dados em invés de uma substituição dos dados.
   1. Antes da remoção: 14.405 clientes
   2. Após a remoção: 14.265 clientes
   3. **Etapa 3 – Transformação das Tabelas em uma tabela**
9. Considerando que as comunicações com status igual a “NAO ENTREGUE” originam de erros no processo de envio do comunicado, gerando um percentual (%) de falhas por diversos motivos (servidor indisponível, número incorreto etc.), como apontado na descrição do case, irei desconsiderar tais comunicações para a análise do problema proposto. Mas para tentar diminuir tais falhas pode-se:
   * analisar o motivo da indisponibilidade do servidor;
   * analisar os possíveis erros no software que gerencia o envio dos comunicados;
   * no caso de dados incorretos do cliente, solicitar uma atualização dos meios de comunicação.
10. Para unir as tabelas “portfolio\_comunicados”, “portfolio\_geral”, “portfolio\_clientes”, e “portfolio\_tpv” em uma tabela, segui a sequência abaixo:
    1. Filtrei a tabela “portfolio\_comunicados” desconsiderando os comunicados com status = “NAO ENTREGUE” e criei a tabela “portfolio\_comunicados\_filtrado”. Isso fez com que, do total de 403.740 comunicados, apenas 212.533 comunicados serão analisados.
    2. Sobre a tabela “portfolio\_geral”, à princípio, selecionei apenas as variáveis mais interessantes e necessárias para a análise. No caso, as variáveis: 'contrato\_id', 'dt\_ref\_portfolio', 'nr\_documento', 'vlr\_pgto\_esperado', 'vlr\_pgto\_realizado', 'vlr\_saldo\_devedor', 'vlr\_saldo\_devedor\_esperado', 'vlr\_desembolsado', 'status\_contrato', 'dsp', 'dspp', 'flag\_transacao'
    3. Com base nos filtros das etapas anteriores, criei uma tabela única (portfolio) partindo da tabela “portfolio\_comunicados\_filtrado” e unindo-a com a tabela “portfolio\_geral” (com as variáveis selecionadas). Foi realizado um merge, através das variáveis 'contrato\_id' e 'dt\_ref\_portfolio' da tabela “portfolio\_comunicados\_filtrado” e as variáveis 'contrato\_id' e 'dt\_ref\_portfolio' da tabela “portfolio\_geral”.
    4. Combinei a tabela única (portfolio), resultado do merge acima, com a tabela “portfolio\_clientes” através da variável 'nr\_documento' de ambas as tabelas.
    5. E por fim, combinei a tabela única (portfolio), resultado do merge acima, com a tabela “portfolio\_tpv” através das variáveis 'nr\_documento' e 'dt\_ref\_portfolio' da tabela “portfolio” e as variáveis 'nr\_documento' e 'dt\_transacao' da tabela “portfolio\_tpv”.
    6. **Etapa 4 – Criação de novas variáveis**
11. Para análise e exploração dos dados, criei três variáveis: “semana\_acao” (que contém os dias da semana), “ano\_acao” (que contém apenas o ano) e “mes\_abrev\_acao” (que contém o nome do mês abreviado). As três variáveis partiram da variável “data\_acao” da tabela final “portfolio”.
    1. **Etapa 5 – Exportação dos dados**
12. Após criar a tabela única (portfolio) com todos os dados necessários, exportei para gerar o primeiro entregável (o arquivo “tabela\_final.csv”).
    1. **Etapa 6 – Análise e Exploração dos dados**
13. Por questão de tempo, a análise e exploração dos dados foram realizadas no processo de desenvolvimento do Dashboard. E o resultado das análises e insights seguem abaixo.
14. Considerando todo o período dos dados (07/01/2020 a 19/04/2022), a quantidade de comunicados entregues, lidos e respondidos por contrato se concentram, em média, no começo da semana (segunda e terça), mas o valor do pagamento realizado se concentra mais na metade da semana para frente (quarta, quinta e sexta) (como pode ser visto na imagem abaixo). A escolha do valor do pagamento realizado foi feita porque acredito que ela indica a efetividade da recuperação dos contratos.

Timeline

Description automatically generated

1. Considerando os anos individualmente, o padrão se mantém, como pode ser visto nas imagens abaixo.

Período: 07/01/2020 a 31/12/2020

Timeline

Description automatically generated

Período: 01/01/2021 a 31/12/2021

Timeline

Description automatically generated

Período: 01/01/2022 a 19/04/2022

A picture containing chart

Description automatically generated

Ou seja, a quantidade de comunicados se concentram no início da semana (segunda e terça), mas o valor do pagamento realizado se concentra mais na metade da semana para frente (quarta, quinta e sexta), período em que a quantidade de comunicados é menor.

1. Sobre o período (Ano e Mês), verifiquei que a quantidade de comunicados por contrato é, em média, de 3 a 4 comunicados por mês. Em 2021, vemos que a concentração de valores de pagamento realizado se concentram em volume no início do mês.

A picture containing timeline

Description automatically generated

1. Analisando a variável “acao”, que identifica o tipo da ação de comunicação utilizada pelo time de cobrança para identificar qual foi a campanha executada, podemos verificar que a “Campanha de Observação”, a primeira campanha acionada pela régua de comunicação, já possui uma eficiência maior que a média. E isso pode ser verificado comparando as duas imagens abaixo:

Todos os tipos de ação e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Campanha de Observação e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Aqui podemos analisar que, considerando apenas a “Campanha de Observação”, a média de comunicados são menores, mas o valor de pagamento realizado é maior:

* Para os comunicados entregues, em torno de 30% maior na quinta e sexta-feira.
* Para os comunicados lidos, em torno de 36% maior na quinta-feira e em torno de 49% maior na sexta-feira.
* E para os comunicados respondidos, em torno de 56% maior na quinta-feira e em torno de 76% maior na sexta-feira.

E um comportamento semelhante (ou seja, maior valor de pagamento realizado nas quintas e sextas-feiras) pode ser observado analisando os anos individualmente.

1. Analisando por região, podemos notar algumas especificidades, como descrito abaixo:

Região Sudeste e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Na região Sudeste, notamos um comportamento semelhante à média total, uma vez que concentra 47,6% do total de comunicados.

Região Centro-Oeste e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Na região Centro-Oeste, também notamos um comportamento semelhante à média total.

Região Nordeste e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Na região Nordeste, apesar da média de comunicados, notamos uma leve queda no valor de pagamento realizado na sexta-feira para os comunicados entregues e lidos. E uma queda mais acentuada para os comunicados respondidos.

Região Norte e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Na região Norte, apesar da média de comunicados, notamos uma queda no valor de pagamento realizado na sexta-feira.

Região Sul e Período entre 07/01/2020 e 19/04/2022

Timeline

Description automatically generated

Na região Sul, notamos uma queda no valor de pagamento realizado a partir de quarta-feira apenas para os comunicados respondidos.

1. **Conclusões e Insights**

Diante das análises dos dados disponibilizados, verifiquei que 305 contratos que deveriam receber algum comunicado para retornar o pagamento ou quitar o saldo devedor não receberam o comunicado.

Além disso, a eficiência do envio das comunicações é de 52,64% (ou seja, dos 403.740 comunicados, apenas 212.533 comunicados são entregues, lidos ou respondidos).

Assim, analisando estes comunicados entregues, lidos ou respondidos, o maior volume médio de valor de pagamento realizado se encontra para um a três comunicados, em média, por semana e comunicados focados nos dias da semana: quarta, quinta e sexta-feira.

Ao longo do ano, verifiquei que, em média, dois a três comunicados são enviados por mês. E o volume médio de valor de pagamento realizado se concentra no final do ano em 2020, e no começo do ano em 2021.

Notei também que o comportamento dos comunicados versus valor do pagamento realizado já na primeira campanha, “Campanha de Observação”, tem uma alta eficiência. E cada região possui suas especificidades.