DATA CHALLENGE 2022

Uma iniciativa Stone

“You must gain control over your money or the lack of it will forever control you” Dave Ramsey

Sumário

Preenchendo o framework de avaliação

Abaixo, você encontrará todas as etapas que deverão ser preenchidas. Preste muita atenção durante o preenchimento e não se esqueça de revisar este documento antes de enviá-lo. É através deste documento que boa parte da avaliação será conduzida **então, capricha! ;)**

Validações iniciais

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

Tabela Geral + TPV + Clientes (Merge): 14754.
Tabela Comunicados: 12175.

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

Tabela Geral: 2 contratos foram excluídos por não terem informações de Cidade/Estado.
Tabela comunicados: Não houve exclusão.

Fluxo de etapas realizadas

0.0. Etapa 0 - IMPORTS

Etapa criada para guardar bibliotecas, funções auxiliares, e datasets importados para a resolução do case.

1.0. Etapa 1 - DESCRIÇÃO E LIMPEZA DOS DADOS

Etapa criada para observar as características e comportamento de cada dataset. Para cada um, o seguinte fluxo foi realizado:

1.1. Descrição das colunas
1.2. Dimensões dos dados
1.3. Checagem de NA e Substituição
1.4. Tipos de Dados e Alteração
1.5. Estatística Descritiva
1.6. Filtragem de dados

Os resultados estão bem detalhados no Notebook “descricao_e_limpeza_dos_dados.ipynb”, aqui trarei os pontos principais da filtragem feita em cada dataset pra não tornar o documento massante.

1. **Portfolio Geral:** Todas as colunas de data foram convertidas para datetime por terem vindo como string. Já a coluna “prazo”, convertida para int por ter vindo como float, mas representar parcelas inteiras.

   Durante a estatística descritiva, foram encontrados alguns problemas. Abordei a solução da seguinte forma:

   - **Valor mínimo e mediana de valor tarifa igual a 0:** Função lambda na coluna valor tarifa pra substituir os 0 por 1% do valor do empréstimo. Aproximadamente 7 milhões de linhas não estavam

com o valor por algum motivo.

Os demais, optei por resolver após o merge da tabela portfolio geral com a de tpv.

Portfolio Clientes: Não foi necessária nenhuma alteração nos tipos de dados. Existiam documentos de clientes repetidos e clientes que não possuíam cidade nem estado registrados. Para resolver a duplicidade dos documentos, dropei os duplicados. Também dropei os clientes que não possuíam cidade/estados pois poderia atrapalhar na análise exploratória poteriormente.

Portfolio Comunicados: Todas as colunas de data foram convertidas para datetime por terem vindo como string. Não teve problemas na estatística descritiva.

Portoflio TPV: Todas as colunas de data foram convertidas para datetime por terem vindo como string. Optei também por resolver os problemas após o merge com a geral.

Portfolio Geral + TPV: Left merge. Cada linha da tabela geral representa um dia corrido de vários contratos e um cliente pode ter mais de um contrato. Cada linha da tabela TPV representa um dia em que foram feitas transações na maquininha Stone do cliente (não corridos, pois se não foi feita transação, não tem registro). A ideia do Left Merge é que pra cada número de documento e data de referência da base geral que bater com o documento e data de transação da base TPV seja adicionada as informações de transação, e os para os que não baterem, NA, demonstrando que não houve movimentação na maquininha no dia.

Para tratar os NA's, dropei dt_transacao da tabela tpv por ser a mesma coisa de dt_ref_portfolio. Preenchi com 0 os valores NA de qtde_transacoes e vlr_tpv, significa que não ocorreu transação naquele dia corrido de contrato

Foram encontrando muitos problemas durante a estatística descritiva, suas soluções foram abordadas da seguinte forma:

- **Percentual de retenção máximo acima de 100%:** Observando os

outliers e fazendo uma pesquisa sobre taxas de retenção decidi correr o risco de tomar todos os valores como erro de digitação e dividi-los por 10, pois representam taxas muito acima do comum. Ex: Antiga taxa 120% (1.2) passa a ser 12% (0.12).

- **Valor mínimo do pagamento realizado negativo:** Drop nas linhas com valor de pagamento realizado negativo, eram desprezíveis em comparação ao tamanho do dataset.
- **Valor do saldo devedor máximo negativo (A stone está devendo pro próprio cliente? Muito estranho):** Drop nas linhas com valor do saldo devedor máximo negativo, eram desprezíveis em comparação ao tamanho do dataset.
- **Verificar se existem dias com flag 1, porém tpv 0 (não faz sentido, já que tpv 0 indica que não teve transação e flag 1 indica que sim):** Atribuição do valor 0 pras flags que estavam como 1 mas no dia o valor transacionado e a quantidade de transações eram iguais a 0. Tomo como premissa desse dataset que o valor do pagamento realizado não tem a ver com a maquininha, e sim como algo que o cliente pague diretamente a stone. O valor pago através da maquinha é descontado de acordo com o percentual de transação e somado ao valor do pagamento realizado, se houver.
- **Quantidade mínima de transações negativa:** Drop nas linhas com a quantidade de transações negativa, eram desprezíveis em comparação ao tamanho do dataset.
- **Valor mínimo de tpv negativo:** Drop nas linhas com valor do tpv negativo, eram desprezíveis em comparação ao tamanho do dataset.
- **Mais de uma linha “Settled”:** Drop em todas as linhas após a primeira data do status do contrato dado como quitado. Isso ajudou a reduzir o tamanho do dataset, visto que pra fins da nossa análise não eram necessárias as linhas depois do contrato ser quitado.

2.0. Etapa 2 - FEATURE ENGINEERING

Etapa criada para derivação de novas features que possam complementar a

Análise Exploratória de Dados.

1. **Portfolio Clientes:**
    a. **Features criadas:**
        i. **Região (Norte, Nordeste, Centro-Oeste, Sudeste, Sul)**
2. **Portfolio Comunicados:**
    a. **Features criadas:**
        i. **Coluna 'negativado' binária onde 1 significa que esse contrato chegou na fase negativado em algum momento e 0 que não.**
3. **Portfolio Geral + TPV:**
    a. **Features criadas:**
        i. **Coluna 'settled' binária onde 1 vai representar que o contrato foi quitado em algum momento, e 0 que nunca foi quitado.**
        ii. **Dia da semana dt_ref_portfolio**
        iii. **Dia dt_ref_portfolio**
        iv. **Semana do ano dt_ref_portfolio**
        v. **Mês dt_ref_portfolio**
        vi. **Trimestre dt_ref_portfolio**
        vii. **Ano dt_ref_portfolio**

3.0. Etapa 3 - ANÁLISE EXPLORATÓRIA DE DADOS

Etapa criada para, depois de ter os datasets devidamente limpos e preparados, analisar o comportamento dos dados em busca de insights para a solução do problema proposto.

Top 3 insights:

1. **A Stone opera em um prejuízo de R$:13,948,731 e esse prejuízo vem de apenas 35% dos contratos.**
2. **80% do valor total dos empréstimos concebidos estão concentrados em 41% dos contratos.**
3. **75% dos contratos voltam a pagar antes de 10 dias.**

2.3. Conclusões e insights

Qual é a curva ideal de vezes que devemos acionar um cliente?

Atualmente, a stone atua num prejuízo de 14 milhões de reais, que vem de apenas 35% dos contratos. Desses, 71% pertencem aos segmentos: Alimentação, Varejo e Bens Duráveis. Foi constatado durante a análise que 75% dos contratos com pagamento atrasado voltam a pagar nos 10 primeiros dias da régua de acionamento. Depois disso, a inadimplência se torna cada vez mais provável.

Tendo esses dados em mente, a minha proposta é dividir a régua de acionamento em duas: Prioridade e Padrão.

A régua de prioridade vai servir para os três setores que mais trazem prejuízo e ser um pouco mais eficiente em sua comunicação, adicionando uma ligação na etapa de parcelamento. Isso traz mais custo para a equipe de comunicação, mas pode reduzir significativamente o prejuízo visto o aumento da taxa de resposta dos clientes, por ser uma chamada de voz.

DSP		
Régua DSP Prioridade		
Campanha	***Regra***	***Canal Acionamento***
Campanha de Observação	5 dias sem pagamento	Email / Whatsapp
Campanha Parcelamento	10 dias sem pagamento	**Ligação** / Whatsapp
Campanha Boleto Quitado	15 dias sem pagamento	Email / Whatsapp
Campanha Pré Negativação	25 dias sem pagamento	Email / Whatsapp
Campanha de Negativação	45 dias sem pagamento	Email / Whatsapp
Campanha Boleto Quitado	70 dias sem pagamento	Email / Whatsapp

DSPP		
Régua DSPP Prioridade		
Campanha	***Regra***	***Canal Acionamento***

Campanha de Observação	10 dias sem pagamento	Email / Whatsapp
Campanha Parcelamento	20 dias sem pagamento	**Ligação** / Whatsapp
Campanha Boleto Quitado	30 dias sem pagamento	Email / Whatsapp

A régua padrão será semelhante a mostrada no ínicio do desafio, porém com um tempo mais curto em relação às campanhas, tendo em vista a relação direta do tempo sem pagar com a diminuição da taxa de retorno. Essa não traz custo adicional algum para a equipe, apenas algumas rápidas mudanças no processo.

DSP		
Régua DSP Padrão		
Campanha	***Regra***	***Canal Acionamento***
Campanha de Observação	5 dias sem pagamento	Email / Whatsapp
Campanha Parcelamento	10 dias sem pagamento	Email / Whatsapp
Campanha Boleto Quitado	15 dias sem pagamento	Email / Whatsapp
Campanha Pré Negativação	25 dias sem pagamento	Email / Whatsapp
Campanha de Negativação	45 dias sem pagamento	Email / Whatsapp
Campanha Boleto Quitado	70 dias sem pagamento	Email / Whatsapp

DSPP		
Régua DSPP Padrão		
Campanha	***Regra***	***Canal Acionamento***
Campanha de Observação	10 dias sem pagamento	Email / Whatsapp

Campanha Parcelamento	20 dias sem pagamento	Email / Whatsapp
Campanha Boleto Quitado	30 dias sem pagamento	Email / Whatsapp

Na minha visão o segredo não é a curva de vezes, mas a de tempo! E também a efetividade do canal de acionamento. No final, o cliente precisa ser alcançado, e nada melhor do que uma ligação para isso.

Desde já, agradeço a equipe Stone pelo excelente hackathon!