vermelho: 39
verde: 206
azul: 90
hexa: #27ce5a

vermelho: 35
verde: 200
azul: 84
hexa: #23c854

vermelho: 22
verde: 180
azul: 69
hexa: #16b445

vermelho: 19
verde: 164
azul: 60
hexa: #13a43c

Stone's Data Science Case Study
Chargeback Investigation

Dev: Mateus Broilo
Pelotas, RS, BR – 24/05/2022

MVP: https://github.com/broilo/stone/

First Things First: A brief introduction

atributos

target

cardinalidade

	Dia	Hora	Valor	Cartão	CBK
0	2015-05-01 00:00:00	00:01:54	36.54	536518******2108	Não
1	2015-05-01 00:00:00	00:03:46	36.54	536518******2108	Não
2	2015-05-01 00:00:00	00:08:50	69	453211******1239	Não
3	2015-05-01 00:00:00	00:27:00	193.43	548827******1705	Não
4	2015-05-01 00:00:00	01:32:46	132	531681******9778	Não
...	...	...	...	...	...
11123	2015-05-30 23:07:01	53	514868******7409	Não	NaN
11124	2015-05-30 23:08:47	15	439354******5281	Não	NaN
11125	2015-05-30 23:15:24	20	549167******1648	Não	NaN
11126	2015-05-30 23:17:41	70	518759******8384	Não	NaN
11127	2015-05-30 23:51:31	20	518759******0329	Não	NaN

11128 rows × 5 columns

Aba 1

problema

	Dia	Hora	Valor	Cartão	CBK
0	2015-06-01	00:02:25	112.00	541555******5965	NaN
1	2015-06-01	00:30:45	112.00	406669******7350	NaN
2	2015-06-01	00:43:20	18.34	541187******4535	NaN
3	2015-06-01	00:46:46	55.00	554927******5629	NaN
4	2015-06-01	00:47:50	50.00	498407******2077	NaN
...	...	...	...	...	...
11815	2015-06-29	23:33:15	161.00	406669******8294	NaN
11816	2015-06-29	23:38:18	103.50	490172******5444	NaN
11817	2015-06-29	23:40:11	60.89	498407******2600	NaN
11818	2015-06-29	23:49:48	46.00	467149******3146	NaN
11819	2015-06-29	23:53:44	77.00	470598******4504	NaN

11820 rows × 5 columns

Aba 2

MVP: https://github.com/broilo/stone/

stone

First Things First: A brief introduction

Ref: ClearSale

Mas o que é *Chargeback?*

"Nome dado ao estorno de um valor lançado no cartão, o chargeback pode causar grandes transtornos às empresas"

Chargeback ➡ **Estorno** → Estorno NÃO é reembolso

Mas o que é Estorno?

É uma *"reversão de pagamento"*

Mas quando acontece?

Quando a cobrança é contestada *"invoice/billing dispute"*

Resumindo:

É um mecanismo de proteção ao consumidor

Que cria problemas para o lojista

MVP: https://github.com/broilo/stone/

stone

First Things First: A brief introduction

Por que é um problema?

Pois sem um sistema de controle o lojista pode ter um rombo financeiro

Prevenção de Fraude

Fraude
Produtos com avarias
Mercadorias não recebidas no prazo
Erros de cobrança
Erros de pagmento
Extravios
Compras não autorizadas pela administradora

Por que é um problema para o lojista?

Pois ao aceitar dados de um cartão de crédito, a loja assumi o risco de uma transação ilegítima

MVP: https://github.com/broilo/stone/

stone

First Things First: A brief introduction

	Dia	Hora	Valor	Cartão	CBK
0	2015-05-01	00:01:54	36.54	536518******2108	Não
1	2015-05-01	00:03:46	36.54	536518******2108	Não
2	2015-05-01	00:08:50	69	453211******1239	Não
3	2015-05-01	00:27:00	193.43	548827******1705	Não
4	2015-05-01	01:32:46	132	531681******9778	Não
...	...	...	...	...	...
11123	2015-05-30	23:07:01	53.0	514868******7409	Não
11124	2015-05-30	23:08:47	15.0	439354******5281	Não
11125	2015-05-30	23:15:24	20.0	549167******1648	Não
11126	2015-05-30	23:17:41	70.0	518759******8384	Não
11127	2015-05-30	23:51:31	20.0	518759******0329	Não

11128 rows × 5 columns

	Dia	Hora	Valor	Cartão	CBK
0	2015-06-01	00:02:25	112.00	541555******5965	NaN
1	2015-06-01	00:30:45	112.00	406669******7350	NaN
2	2015-06-01	00:43:20	18.34	541187******4535	NaN
3	2015-06-01	00:46:46	55.00	554927******5629	NaN
4	2015-06-01	00:47:50	50.00	498407******2077	NaN
...	...	...	...	...	...
11815	2015-06-29	23:33:15	161.00	406669******8294	NaN
11816	2015-06-29	23:38:18	103.50	490172******5444	NaN
11817	2015-06-29	23:40:11	60.89	498407******2600	NaN
11818	2015-06-29	23:49:48	46.00	467149******3146	NaN
11819	2015-06-29	23:53:44	77.00	470598******4504	NaN

11820 rows × 5 columns

Conclusions

1. df_sheet1: represents the train/test dataset
   - There're some minor corrections associated with the data integrity/structure that needs to be properly addressed
2. df_sheet2: represenst the validation dataset

Done!

MVP: https://github.com/broilo/stone/

stone

Analytical Record

	COLUMN_NAME	COLUMN_DTYPE	#_NULL	#_NON_NULL	%_NULL	%_NON_NULL	UNIQUE_VALUES
0	Dia	object	0	11128	0.0	100.0	30
1	Hora	object	0	11128	0.0	100.0	10044
2	Valor	float64	0	11128	0.0	100.0	511
3	Cartão	object	0	11128	0.0	100.0	9260
4	CBK	object	0	11128	0.0	100.0	2

	COLUMN_NAME	COLUMN_DTYPE	UNIQUE_VALUES
0	Dia	object	[2015-05-01, 2015-05-02, 2015-05-03, 2015-05-0...
1	Hora	object	[00:01:54, 00:03:46, 00:08:50, 00:27:00, 01:32...
2	Valor	float64	[36.54, 69.0, 193.43, 132.0, 161.0, 110.0, 159...
3	Cartão	object	[536518******2108, 453211******1239, 548827***...
4	CBK	object	[Não, Sim]

Duplicados

	Dia	Hora	Valor	Cartão	CBK
6104	2015-05-15	23:00:20	264.00	515894******6461	Sim
11004	2015-05-30	14:32:17	15.00	514945******7580	Não
11005	2015-05-30	14:32:37	96.42	498408******2729	Não
11006	2015-05-30	14:33:03	35.00	441524******8556	Não
11007	2015-05-30	14:35:14	99.00	546451******1223	Não
...	...	...	...	...	...
11123	2015-05-30	23:07:01	53.00	514868******7409	Não
11124	2015-05-30	23:08:47	15.00	439354******5281	Não
11125	2015-05-30	23:15:24	20.00	549167******1648	Não
11126	2015-05-30	23:17:41	70.00	518759******8384	Não
11127	2015-05-30	23:51:31	20.00	518759******0329	Não

124 rows × 5 columns

Problema de integridade

```
Problem lines: [7779]
```

	Dia	Hora	Valor	Cartão	CBK	Data
7779	2015-05-22	1899-12-30 00:00:00	23.0	498453******6960	Não	2015-05-22 1899-12-30 00:00:00

MVP: https://github.com/broilo/stone/

stone

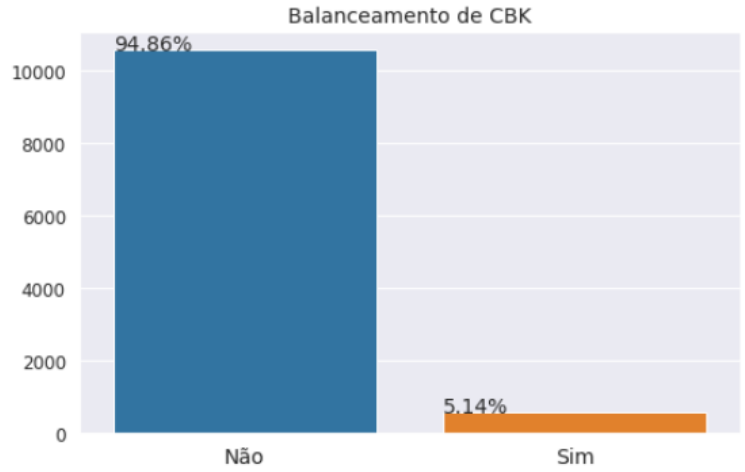
Análises

Qual o comportamento transacional do cliente*?*

Qual o perfil das transações que retornam *chargeback?*

Levantamento de Hipóteses

Group	#	Hypothesis	Result
Valor da transação	**H1**	Transações de maior valor tendem a ter mais estornos que as de menor valor.	Accepted
Dia da transação	**H2**	O estorno é influenciado pelo dia da semana em que a transação é feita.	Accepted
	H3	Transações realizadas em diferentes períodos do mês influenciam no estorno.	Accepted
Horário da transação	**H4**	O horário em que a transação é realizada influência no estorno.	Accepted



- Creating resources
  - Concat Dia and Hora resources and check its length
  - Creating day_of_name, day_of_month and pure_time
- Exploratory Data Analysis
  - Simple Hypotheses
  - Univariate Analysis
    - Resource: Valor
    - Resource: pure_time
    - Resource: day_of_month
    - Resource: day_of_week
    - Target Analysis: CBK
  - Bivariate Analysis
    - Valor VS target
    - day_name VS target
    - day_of_month VS target
    - pure_time VS target
  - Hypothesis Conclusion
  - Multivariate Analysis
    - Valor and pure_time per day_name VS target
    - Valor and pure_time per day_of_month VS target
- Análise Comportamental
  - Recorrência de transação diária por Cartão
  - Recorrência de transação diária de mesmo Valor por Cartão
  - Merge
  - Tempo entre transações do mesmo Cartão num mesmo dia
  - Ticket Médio
    - Convertido
    - Não convertido

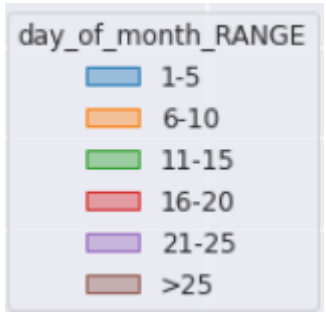
MVP: https://github.com/broilo/stone/

Análises

Cartão	Dia	Hora	Valor	day_of_week	day_name	day_of_month	pure_time	CBK	day_of_month_RANGE	same_day_count	same_day_valor_count	rank_same_day	diff_time
400217******1137	2015-05-06	09:37:46	198.0	2	Wednesday	6	9.629444	Não	6-10	1	1	1.0	0.000000
400217******1353	2015-05-27	23:37:20	172.5	2	Wednesday	27	23.622222	Sim	>25	8	8	1.0	0.000000
400217******1353	2015-05-27	23:38:58	172.5	2	Wednesday	27	23.649444	Sim	>25	8	8	2.0	0.027222
400217******1353	2015-05-27	23:40:15	172.5	2	Wednesday	27	23.670833	Sim	>25	8	8	3.0	0.021389
400217******1353	2015-05-27	23:41:38	172.5	2	Wednesday	27	23.693889	Sim	>25	8	8	4.0	0.023056
...	...	...	...	...	...	...	...	...	...	...	...	...	...

pure_time: número representando o tempo em horas

day_of_month_RANGE: segmentação dos dias mensais



same_day_count: recorrência de transação diária por Cartão

same_day_count: recorrência de transação diária de mesmo valor por Cartão

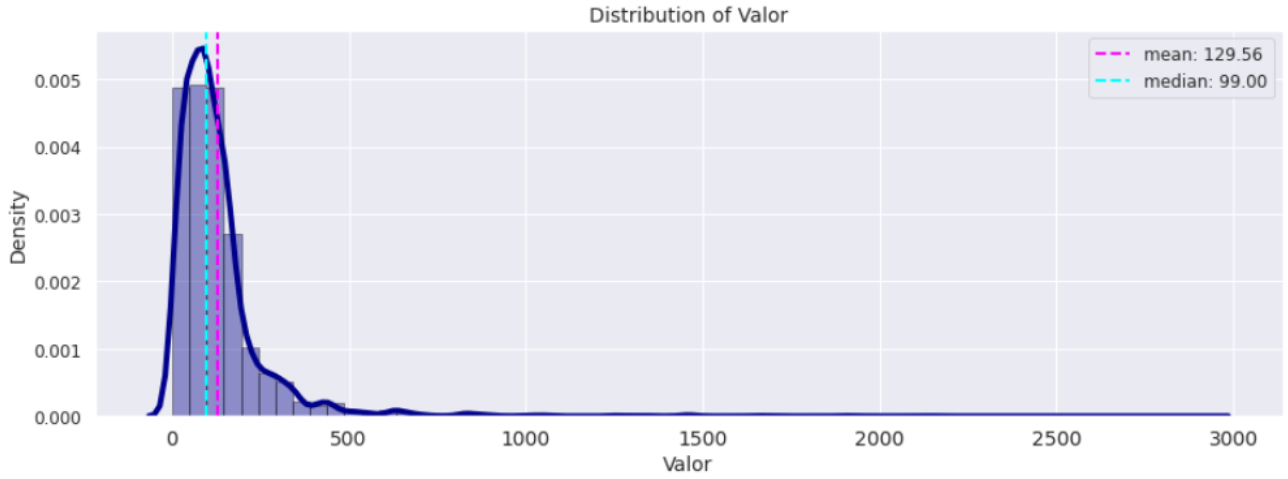
rank_same_day: ordenamento de transação diária por Cartão

diff_time: tempo entre transações do mesmo Cartão num mesmo dia

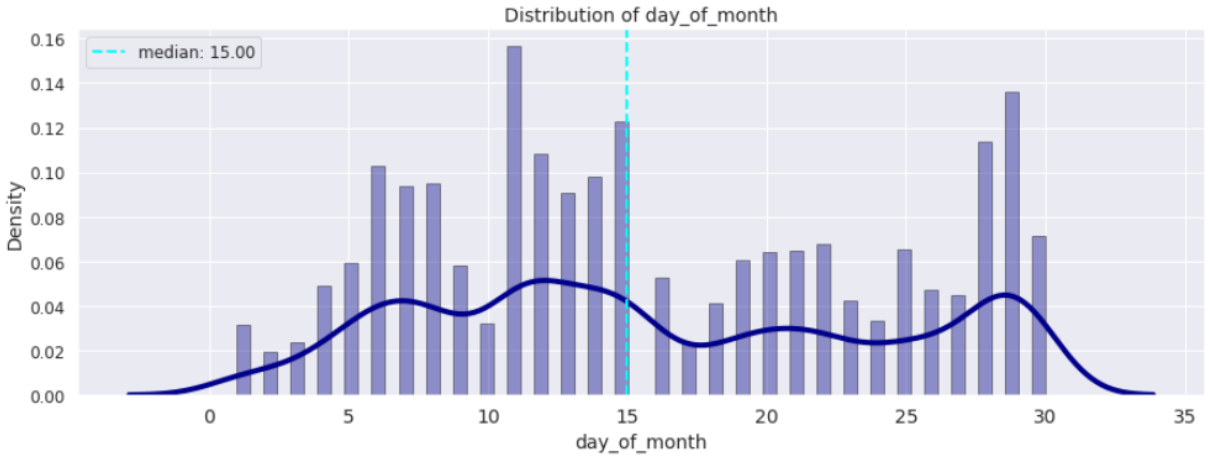
MVP: https://github.com/broilo/stone/

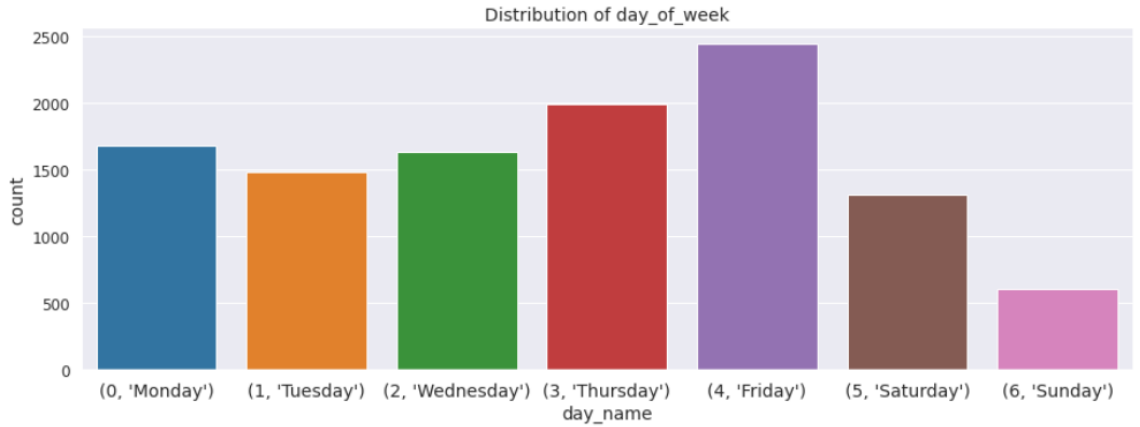
Análises

Qual o comportamento transacional do cliente?



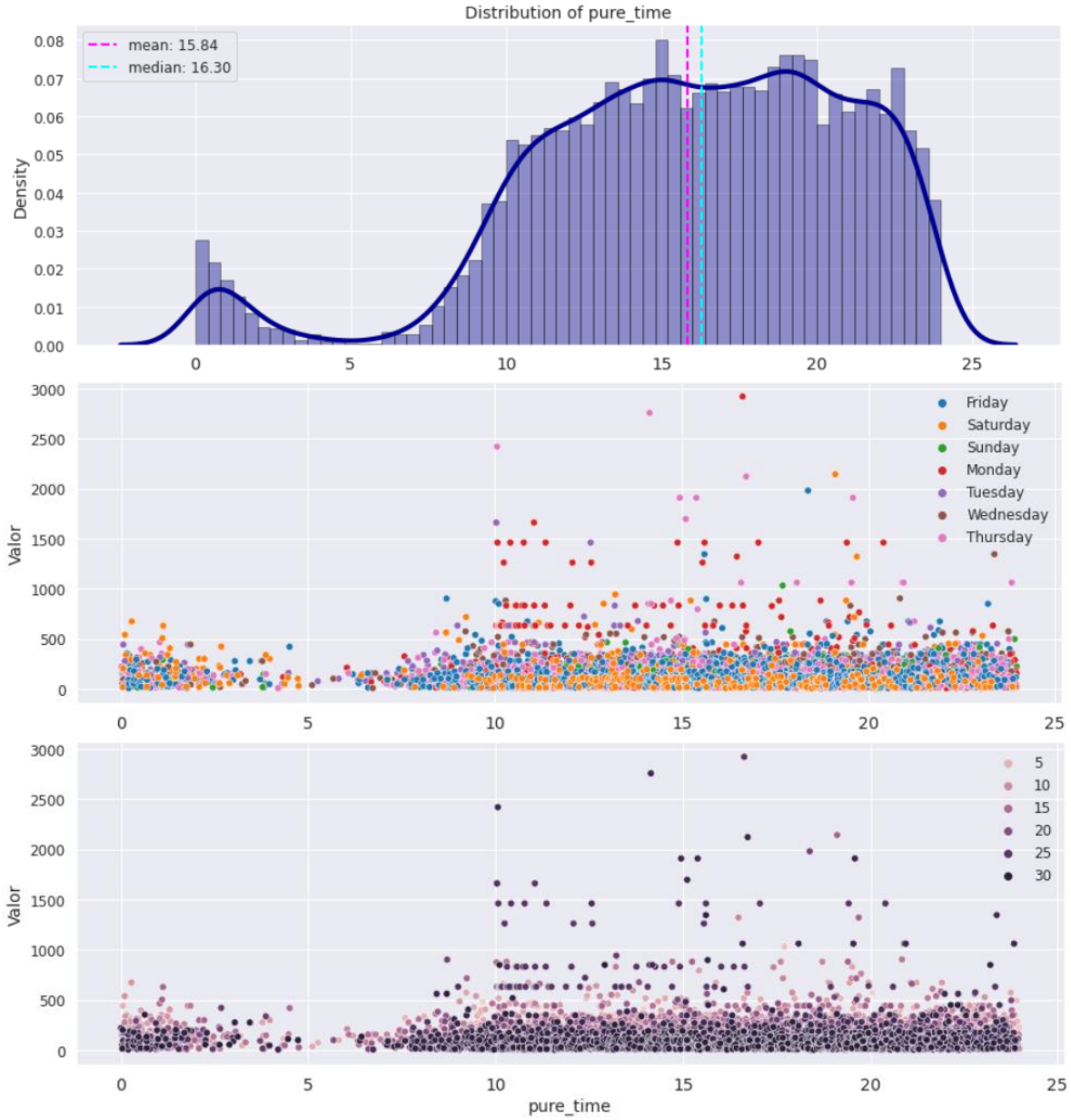
day_name	parque	freqüência	valor (R$ Mil)
Monday	1677	15.1%	275.27
Tuesday	1478	13.3%	179.74
Wednesday	1628	14.6%	209.88
Thursday	1990	17.9%	245.32
Friday	2439	21.9%	307.76
Saturday	1313	11.8%	153.90
Sunday	602	5.4%	69.74





MVP: https://github.com/broilo/stone/

Qual o comportamento transacional do cliente?



Ticket Médio

Convertido: **Não Estorno** -> **R$ 126.65**

Não Convertido: **Sim Estorno** -> **R$ 183.30**

MVP: https://github.com/broilo/stone/

Qual o comportamento transacional do cliente?

Como funcionam as **taxas** da **Stone** Mais

A **taxa** de vendas no débito é de 1,99%. Já para crédito, com recebimento em 1 dia, a **taxa** é de 4,98%. A cada parcela, existe um acréscimo de 1,99%. Para recebimentos em 14 dias, a **taxa** é de 3,98%. 6 de set. de 2019

Ticket Médio

Convertido: **Não Estorno** -> **R$ 126.65**

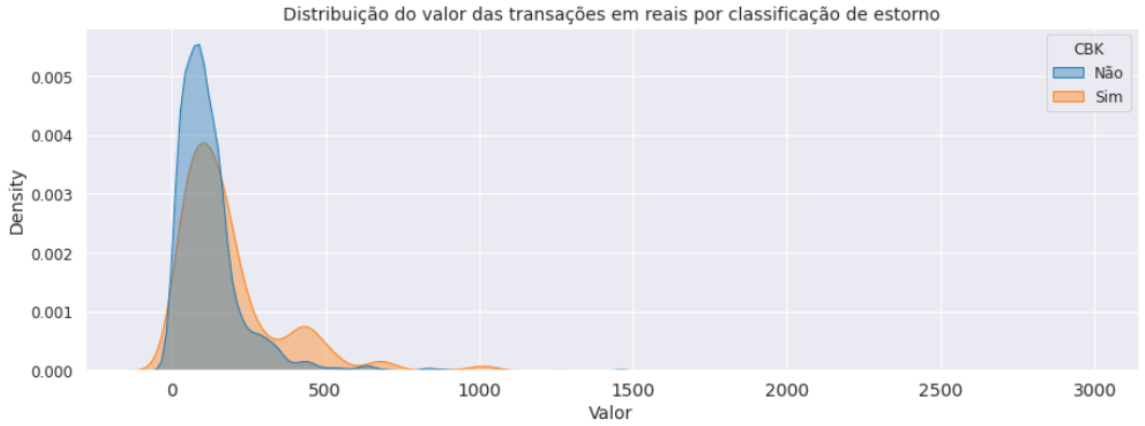
Não Convertido: **Sim Estorno** -> **R$ 183.30**

	CBK	receita (R$ Mi)	parque	ticket médio (R$)	taxa	ticket médio líquido (R$)	PER por transação (R$)
Convertido	Não	1.34	10555	126.65	4.98%	120.34	6.31
Não Convertido	Sim	0.10	572	183.30	4.98%	174.17	9.13

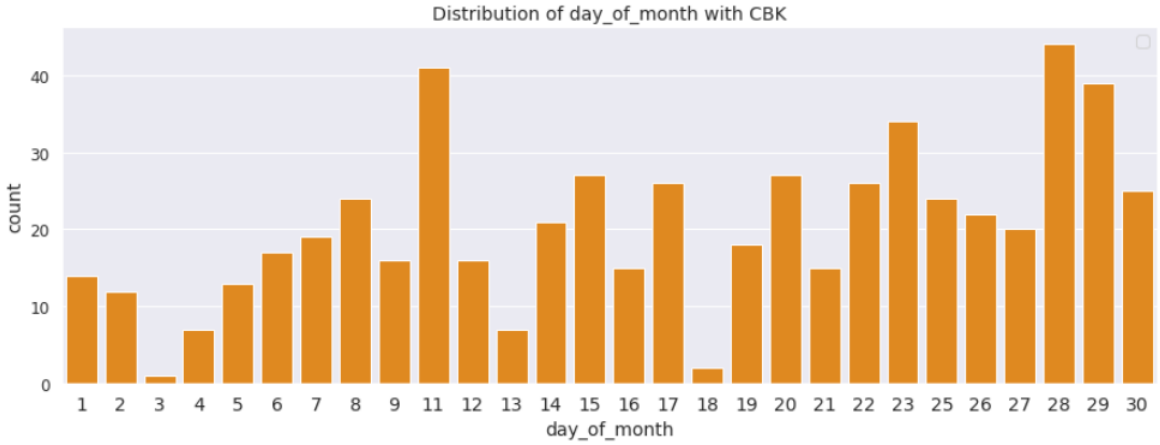
MVP: https://github.com/broilo/stone/

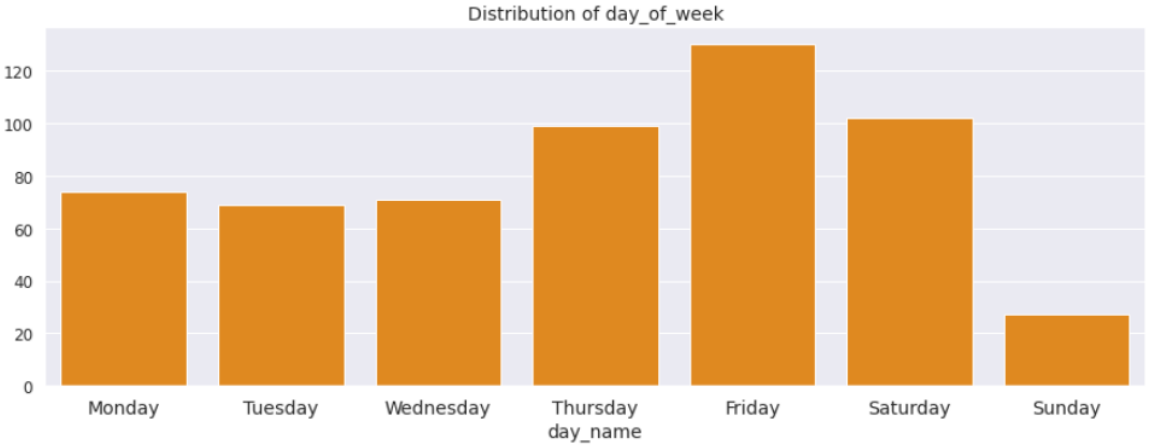
Análises

Qual o perfil das transações que retornam *chargeback?*



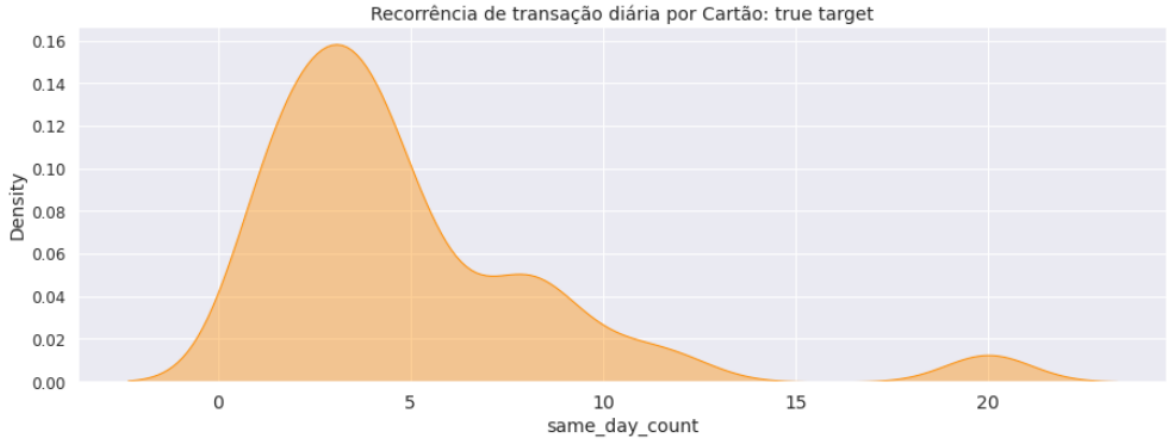
day_name	parque	freqüência	valor (R$ Mil)
Monday	74	12.9%	13.01
Tuesday	69	12.1%	10.74
Wednesday	71	12.4%	18.82
Thursday	99	17.3%	13.94
Friday	130	22.7%	18.56
Saturday	102	17.8%	23.49
Sunday	27	4.7%	6.28

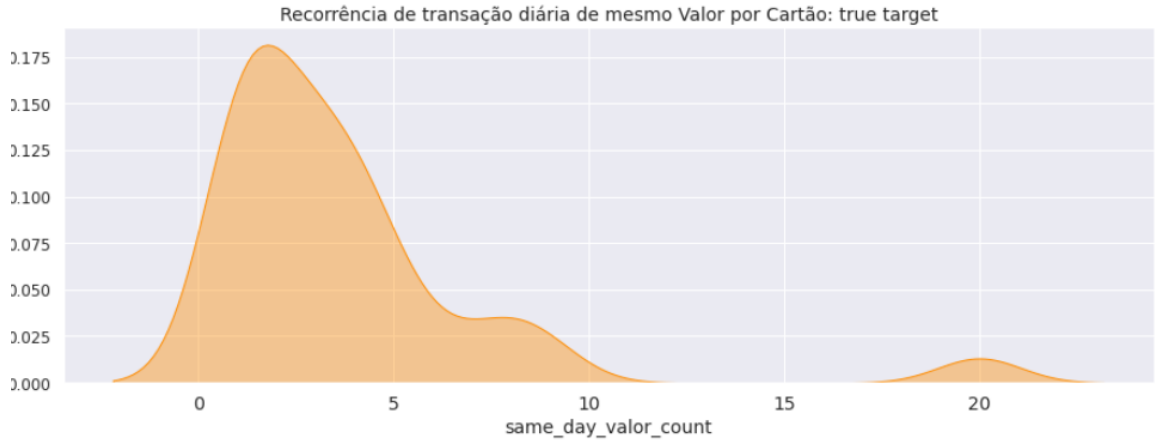


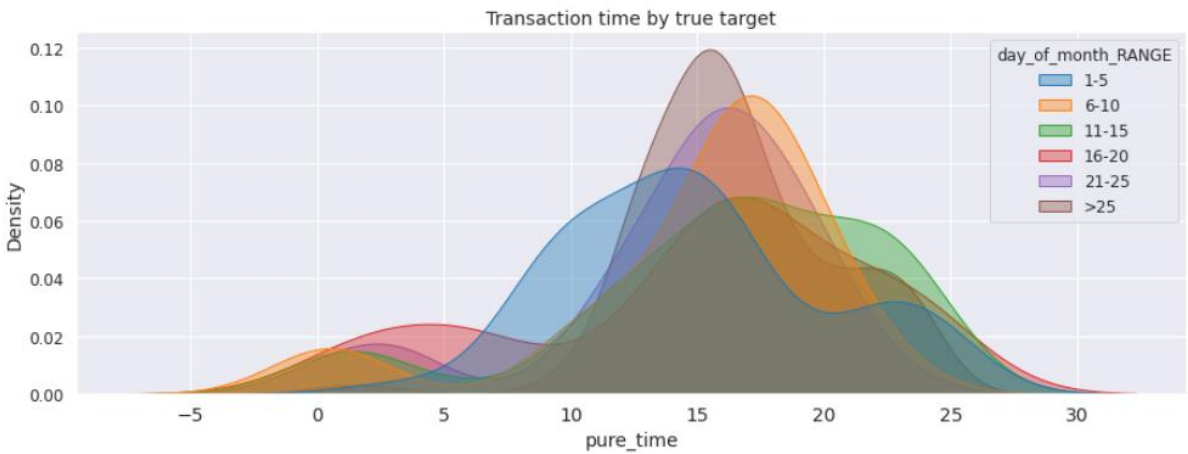


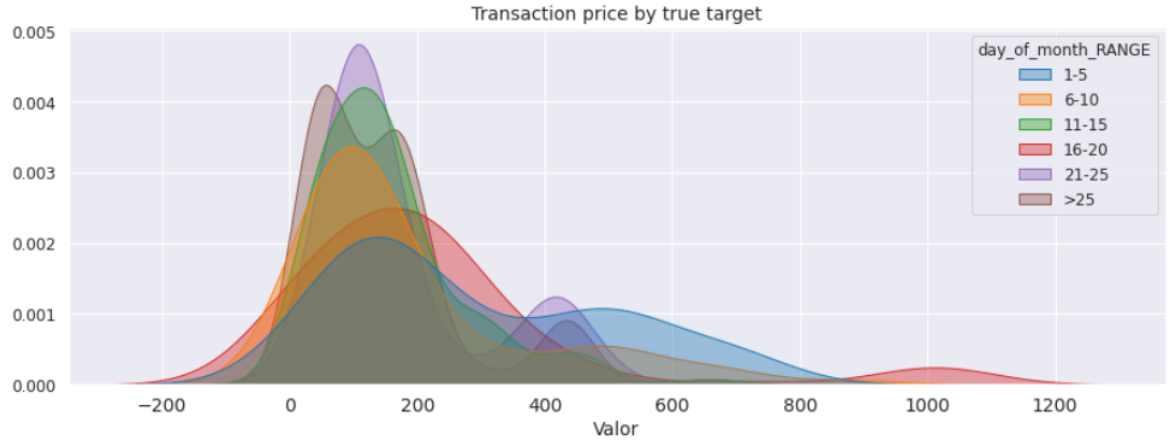
MVP: https://github.com/broilo/stone/

Qual o perfil das transações que retornam *chargeback?*









MVP: https://github.com/broilo/stone/

Modelagem

- Model
  - Splitting Train/Test data: 80/20
    - Normalization
  - Logistic Regression
  - Support Vector Machine
  - XGBoost: Default

POC: Algoritmos Default

Set 1:
Valor
pure_time
same_day_count
same_day_valor_count
rank_same_day
diff_time
day_name
day_of_month_RANGE

Set 2:
Valor
pure_time
rank_same_day
diff_time
day_name
day_of_month_RANGE

Set 3:
Valor
pure_time
same_day_count
same_day_valor_count
rank_same_day
diff_time

Features	Model	Precision	Recall
Set1	Logistic Regression	0.766234	0.572816
	SVC	0.859155	0.592233
	XGBClassifier	0.822917	0.766990
Set2	Logistic Regression	0.760870	0.339806
	SVC	0.869565	0.388350
	XGBClassifier	0.800000	0.582524
Set3	Logistic Regression	0.770270	0.553398
	SVC	0.787500	0.611650
	XGBClassifier	0.865169	0.747573

MVP: https://github.com/broilo/stone/

Modelagem

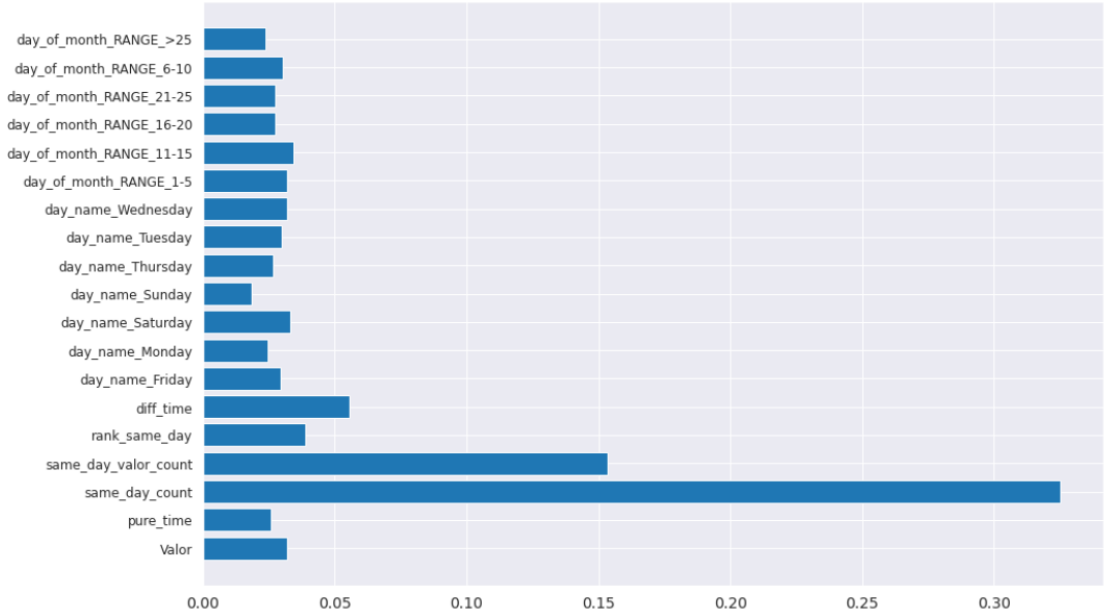
- Splitting Train/Test data: 80/20
- XGBoost
  - Fine-Tuning
    - Best Parameters
    - Feature Importance

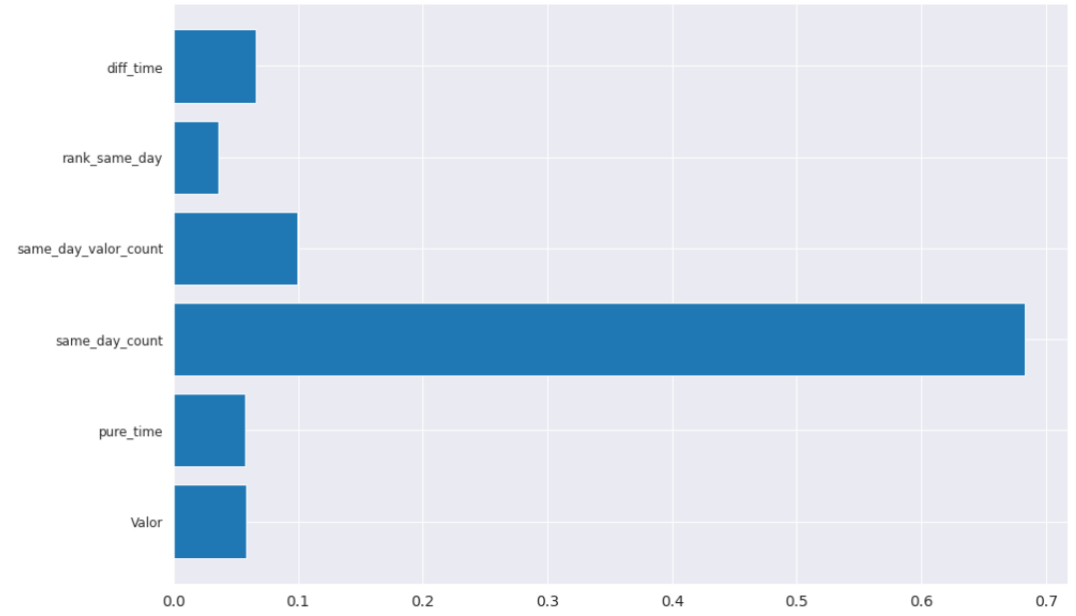
XGBoost Classifier

Set 1:
Valor
pure_time
same_day_count
same_day_valor_count
rank_same_day
diff_time
day_name
day_of_month_RANGE

Set 3:
Valor
pure_time
same_day_count
same_day_valor_count
rank_same_day
diff_time

Features	Model	Precision	Recall
Set1	**XGBClassifier**	0.822917	0.766990
Set3	**XGBClassifier**	0.806452	0.728155





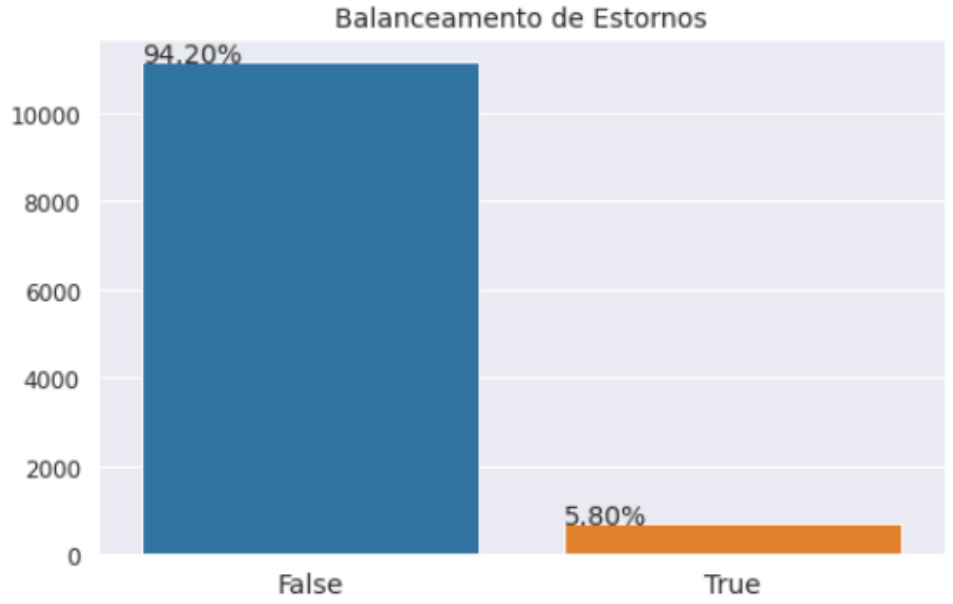
MVP: https://github.com/broilo/stone/

MVP

XGBoost Classifier

Cartão	Data	Previsão	Probabilidade de CBK
400217******0059	03/06/2015 18:44	False	0.76%
400217******3671	02/06/2015 22:01	False	0.09%
400217******3689	16/06/2015 18:18	False	0.92%
400217******3922	15/06/2015 17:17	False	2.00%
400217******4160	11/06/2015 15:26	False	0.05%
400217******5122	23/06/2015 14:40	False	0.05%
400217******6763	02/06/2015 19:56	True	66.88%
400217******6763	02/06/2015 19:57	True	95.59%
400217******6763	02/06/2015 19:58	True	92.76%

- Function's Definition
  - Class of Preprocess Funtions
- Loading data
- Data Preprocessing
  - Date Resources: day_of_week, day_name, day_of_month, pure_time, day_of_month_RANGE
  - Transactions Resources: same_day_count, same_day_valor_count, rank_same_day, diff_time
- Selecting Possible Useful Resources
- Saving preprocessed dataset



	CBK	receita (R$ Mi)	parque	ticket médio (R$)	taxa	ticket médio líquido (R$)	PER por transação (R$)
Convertido	Não	1.26	11135	112.81	4.98%	107.19	5.62
Não Convertido	Sim	0.10	685	146.96	4.98%	139.64	7.32

MVP: https://github.com/broilo/stone/

Regras de Negócio e Impactos

“O que temos”

	CBK	receita (R$ Mi)	parque	ticket médio (R$)	taxa	ticket médio líquido (R$)	PER por transação (R$)	Evasão total (R$)
Convertido	Não	1.26	11135	112.81	4.98%	107.19	5.62	-
Não Convertido	Sim	0.10	685	146.96	4.98%	139.64	7.32	5013.11

Prospectando “O que poderia ser”: corte seco no número de recorrências diárias por transação

	count	mean	std	min	25%	50%	75%	max
rank_same_day	685.0	6.893431	7.231674	1.0	2.0	4.0	9.0	46.0

Considerando rank_same_day .LE. 4

Implica numa redução de 50% de CBK

	CBK	receita (R$ Mi)	parque	ticket médio (R$)	taxa	ticket médio líquido (R$)	PER por transação (R$)	Evasão total (R$)
Convertido	Não	1.26	11135	112.81	4.98%	107.19	5.62	-
Não Convertido	Sim	0.06	368	173.33	4.98%	164.70	8.63	3176.47

Ou seja: Uma redução de 47% da Evasão Total

MVP: https://github.com/broilo/stone/

stone