data® expert.

Data Science & Machine Learning

DS06

PROJETO PRÁTICO

Grupo 18

Grupo 18

Henrique Borgo
55833

Jessica Tizziani
42426

Uákiti Pires
72450

Tópicos:

01

Definições do Projeto

02

Business Understanding

03

Data Understanding

04

Data Preparation

05

Modeling

06

Evaluation & Deployment

01

Definições do Projeto

Empresa: Prepi

Responsáveis:

- Ramon Pereira (CTO e cofundador)
- Dyogo Machado (Head of Aquisitions e Growth)

Uma "**one-stop-shop**" das empreendedoras latino americanas com foco em vendas online.

A visão da empresa é impactar 1 milhão de lojistas até 2026, além de ser referência em Social Commerce na região Latam, mirando no crescimento das vendas digitais.

https://prepi.com.br

Problema de Negócio

Determinação dos Leads com maior probabilidade de converterem em Clientes (Lead Scoring)

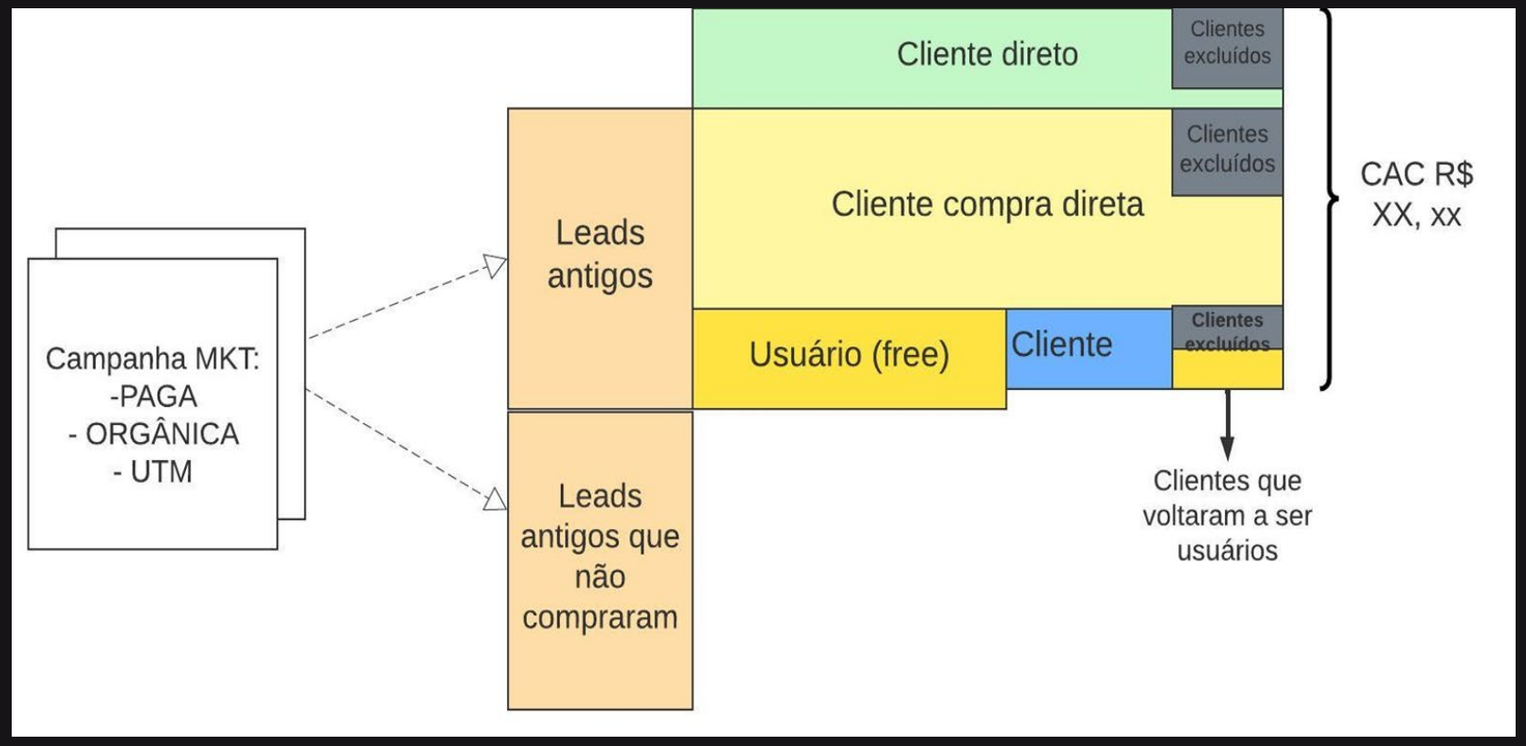
Business Understanding

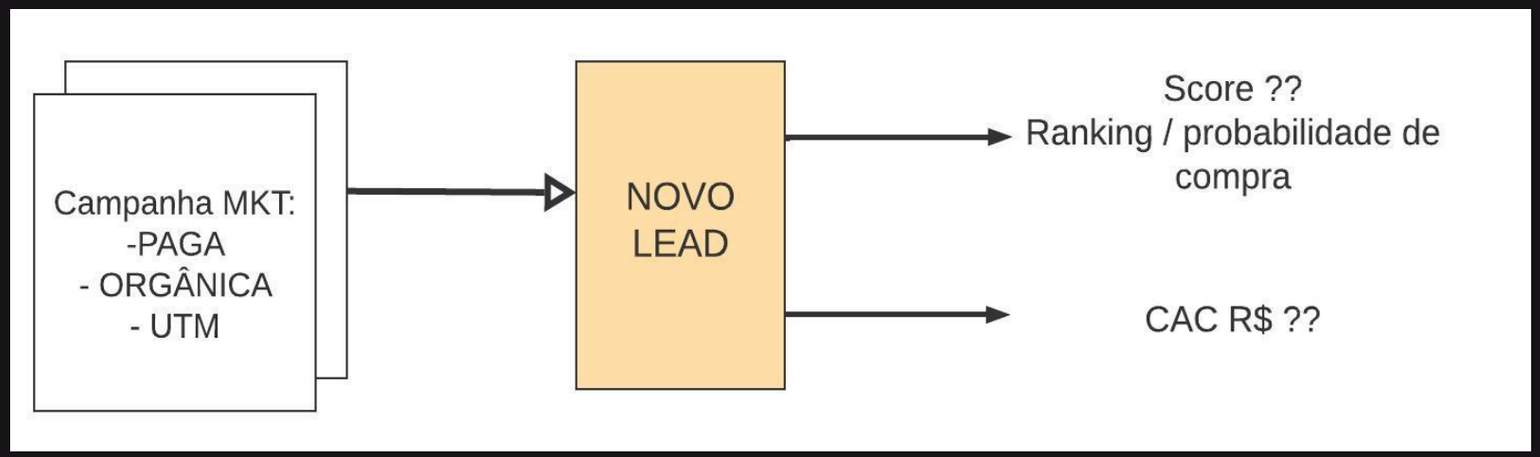
A premissa do negócio é de, com o projeto em questão, encontrar formas de **reduzir custos** ou **elevar o faturamento** e, consequentemente, melhorar o lucro da empresa.

O objetivo principal do projeto é reduzir o **custo por aquisição** (CAC) através da criação de um modelo que indique os clientes ideais (prospects), a partir da categorização da probabilidade destes se converterem em clientes e do seu potencial de serem clientes pagantes, por meio da base de contatos (Leads).

Data Understanding

Análise do funil de vendas:





Data Understanding

Foi disponibilizado banco de dados do tipo .csv, contendo mais de 34.000 linhas e 76 colunas.

A base de dados é resultante da união de 3 diferentes planilhas, sendo: Lista de Leads, Lista de Clientes e Usuários e Custo da Campanhas de Anúncios Pagos.

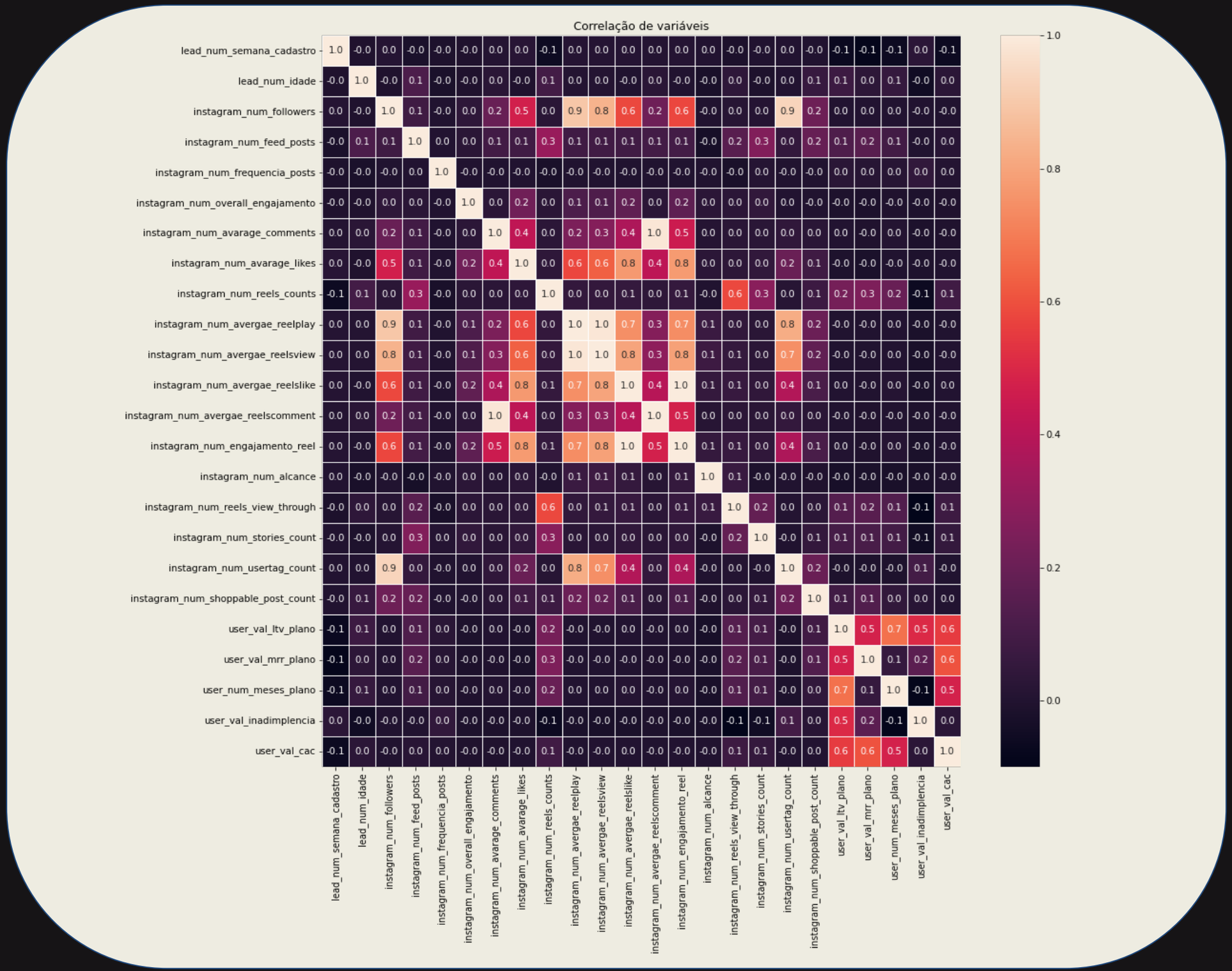
Disponibilizado também os metadados, que foram validados e reformulados.

Data Understanding

Campo	Campo Renomeada	Descrição	Base	Anonimizada
Shoppable Post Count	instagram_num_shoppable_post_count	Quantidade de postagens que possui as ferramentas de Shopping ativas	Instagram	
Public Whatsapp	instagram_hash_public_whatsapp	Número de Whatsapp público no perfil	Instagram	Sim
Public Email	instagram_hash_public_email	Email público no perfil	Instagram	Sim
Public Zip	instagram_public_zip	CEP público no perfil	Instagram	
Category Name	instagram_category_bussiness	Categoria de negócio do Perfil	Instagram	
status	user_status	Status da assinatura na Prepi	Clientes e Usuários	
instagram	user_hash_instagram	Instagram cadastrado durante a criação da Loja virtual	Clientes e Usuários	Sim
Whatsapp	user_hash_whatsapp	Whatsapp cadastrado durante a criação da Loja virtual	Clientes e Usuários	Sim
plano	user_plano	Nome do Plano escolhido	Clientes e Usuários	
Payment	user_payment	Método de pagamento escolhido	Clientes e Usuários	
ltv	user_val_ltv_plano	Valor total da assinatura do plano	Clientes e Usuários	
MRR	user_val_mrr_plano	Valor mensal do plano. LTV / periodicidade	Clientes e Usuários	
periodicidade	user_num_meses_plano	Período de tempo, em meses, da assinatura	Clientes e Usuários	
data_trial	user_dt_trial	Data em que tornou-se testador da Prepi	Clientes e Usuários	
data_cliente	user_dt_cliente	Data em que, de fato, tornou-se cliente da Prepi	Clientes e Usuários	
data_churn	user_dt_churn_prevista	Data prevista para churn da Prepi	Clientes e Usuários	
valor_indimplencia	user_val_inadimplencia	Soma de toda a inadimplência	Clientes e Usuários	
older_indamplencia_date	user_dt_inadimplencia_inicial	Inadimplência mais antiga, ou seja, primeira fatura em débito	Clientes e Usuários	
lead at	user_dt_lead	Data em que se cadastrou e tornouse Lead	Clientes e Usuários	
cidade da loja	user_cidade	Cidade da Loja	Clientes e Usuários	
estado da loja	user_estado	Estado da Loja	Clientes e Usuários	
P1	user_p1	Resposta da primeira pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P2	user_p2	Resposta da segunda pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P3	user_p3	Resposta da terceira pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P4	user_p4	Resposta da quarta pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P5	user_p5	Resposta da quinta pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
Diagnostico	user_diagnostico	Estado atual da Loja	Clientes e Usuários	
Diagnostic Date	user_dt_diagnostico	Data que o diagnóstico foi feito	Clientes e Usuários	
cost	user_val_cac	Custo Prepi para adquirir o Lead/Cliente/Usuário	Clientes e Usuários	

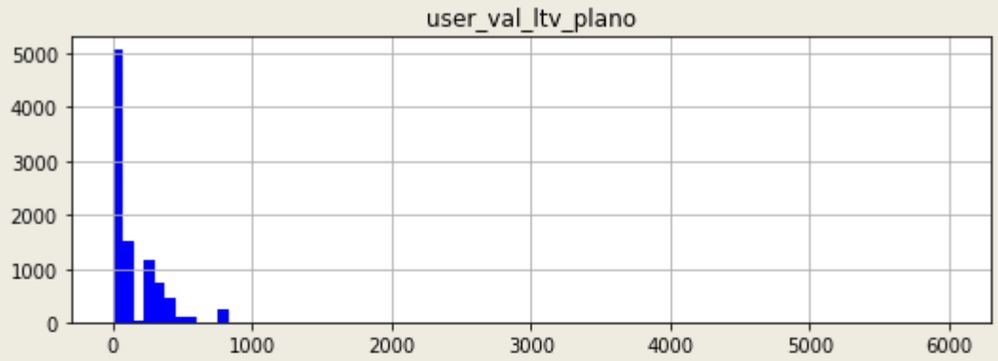
Amostra do dicionário de dados.

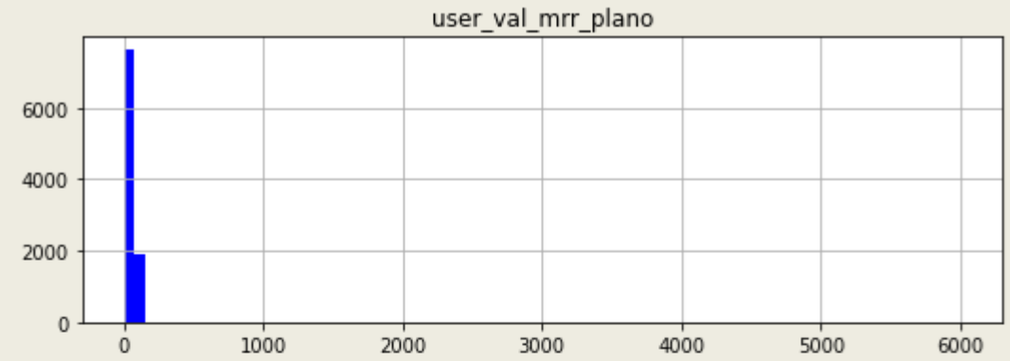
Data Understanding

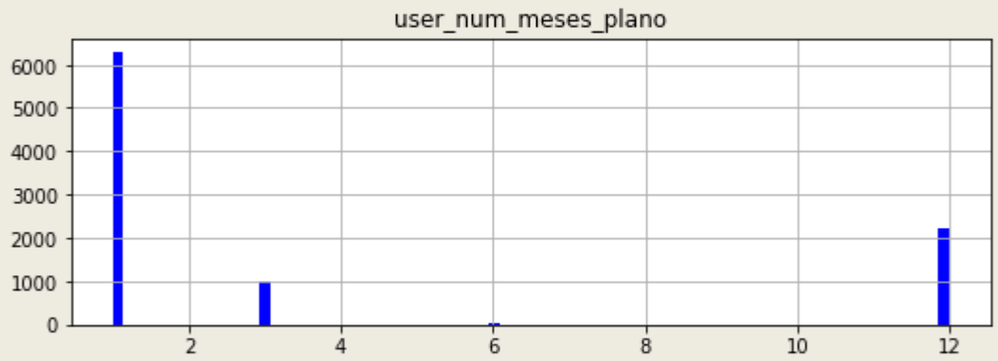


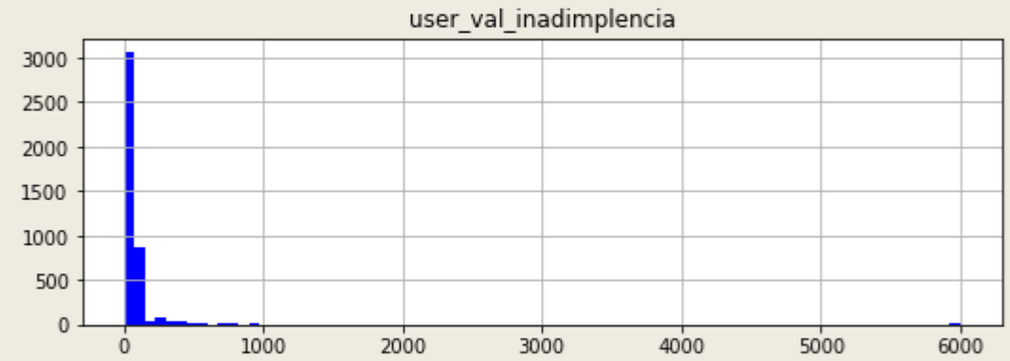
Análise de Correlação entre Variáveis.

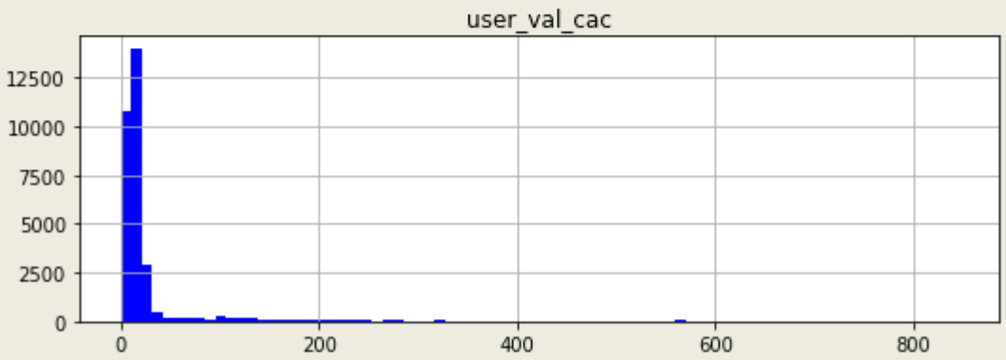
Data Understanding





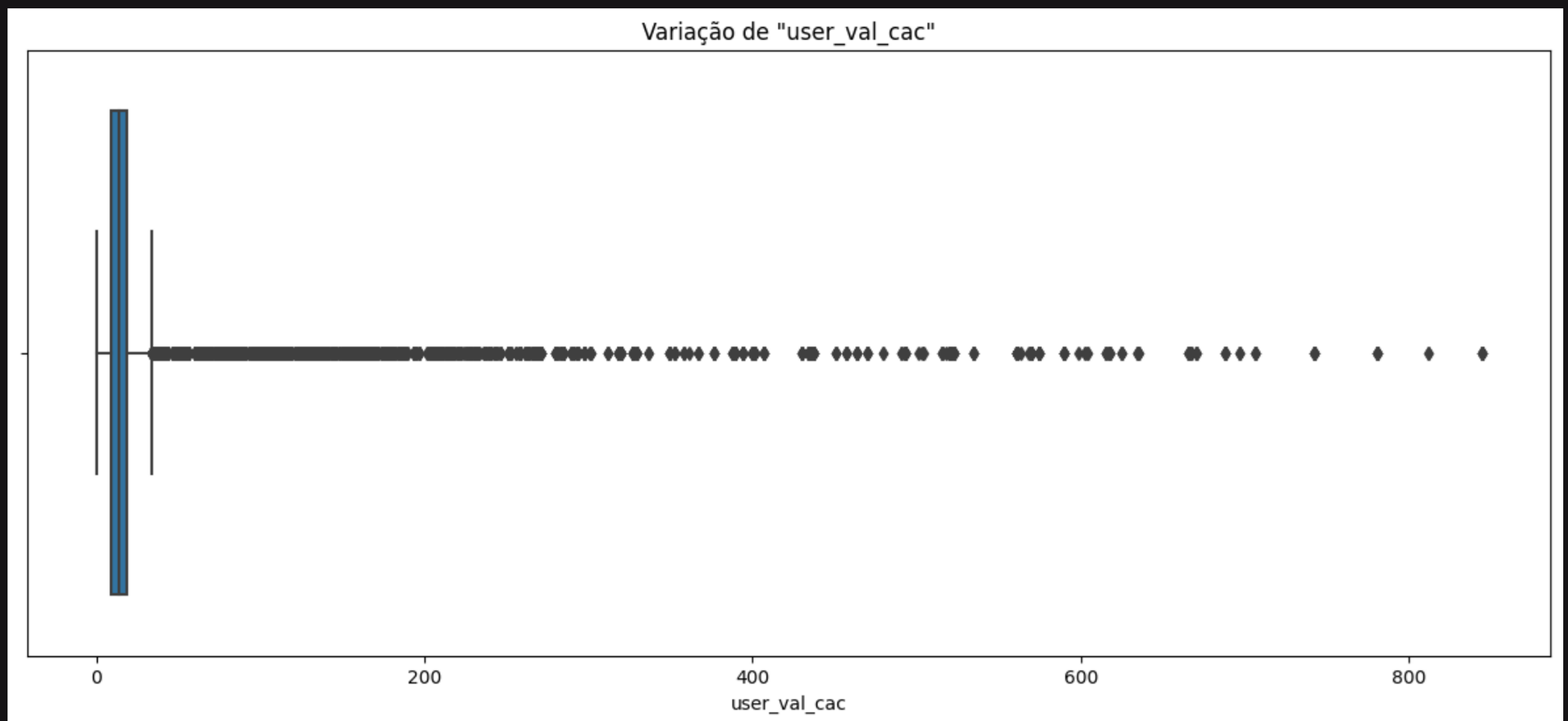






Análise de Distribuição dos Dados (Histograma)

Data Understanding



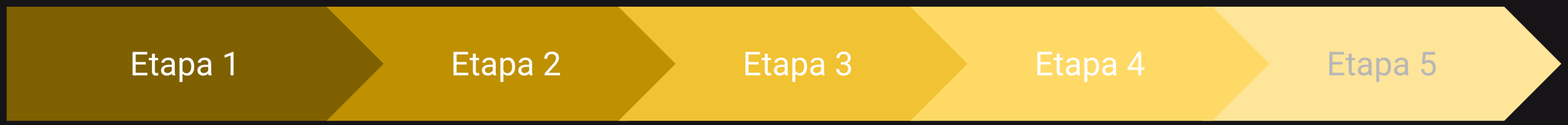

```
df['user_val_cac'].describe()

count    30707.000000
mean        27.378166
std         64.616487
min          0.000000
25%          8.710000
50%         13.550000
75%         18.670000
max        844.140000
Name: user_val_cac, dtype: float64
```

user_status	sum	max	min	mean	median
active	107620.82	844.14	0.00	189.140281	145.870
churned	140010.49	844.14	31.07	184.467049	135.690
debtor	6708.00	697.03	35.85	258.000000	193.350
debtor_trial	3927.71	604.23	118.31	261.847333	293.590
deleted	15495.26	706.02	54.01	218.243099	174.310
suspended	40275.76	844.14	32.21	200.376915	136.670
trial_churned	134160.28	812.12	19.11	214.656448	159.860
user	31572.54	125.13	0.00	7.123768	6.305
waiting_payment	236.62	118.31	118.31	118.310000	118.310

Análise de outliers.

Data Preparation



Selecionar os dados e variáveis com maior relevância

Limpeza dos dados

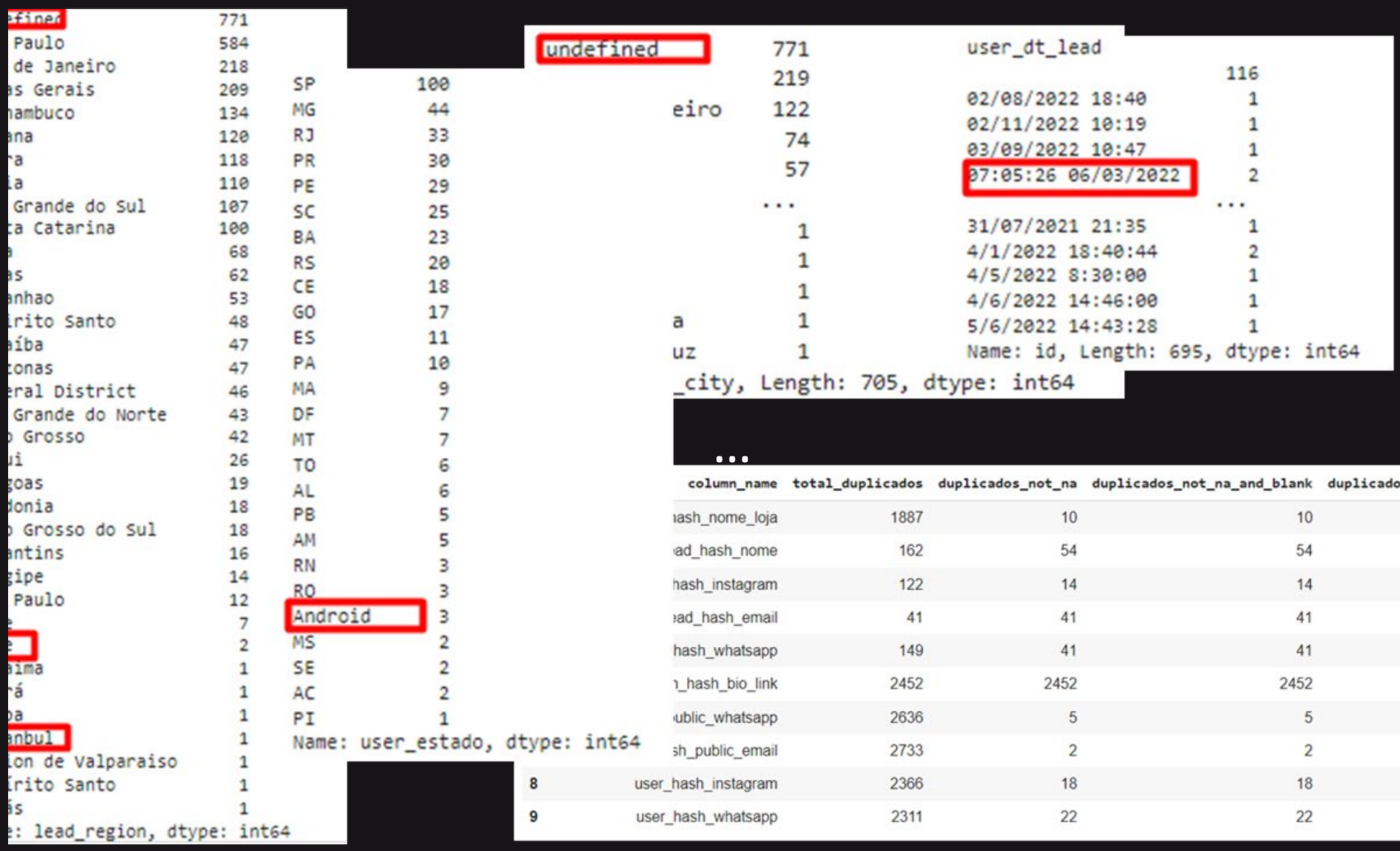
Construção dos dados

Integração dos dados

Formatação dos dados

Data Preparation

Identificado várias colunas com algumas inconsistências e outras com grande quantidade de valores faltantes.



```
finer          771
 Paulo         584
 de Janeiro    218
s Gerais       209
hambuco        134
ana            120
ra             118
ia             110
 Grande do Sul 107
ta Catarina    100
a               68
as              62
anhao           53
irito Santo     48
aíba            47
tonas           47
eral District   46
 Grande do Norte 43
 Grosso        42
ui              26
goas            19
donia           18
 Grosso do Sul 18
antins          16
gipe            14
 Paulo          12
e                7
e                2
aima             1
rá               1
pa               1
anbul            1
ion de Valparaiso 1
irito Santo      1
ás               1
e: lead_region, dtype: int64
```

```
SP    100
MG     44
RJ     33
PR     30
PE     29
SC     25
BA     23
RS     20
CE     18
GO     17
ES     11
PA     10
MA      9
DF      7
MT      7
TO      6
AL      6
PB      5
AM      5
RN      3
RO      3
Android 3
MS      2
SE      2
AC      2
PI      1
Name: user_estado, dtype: int64
```

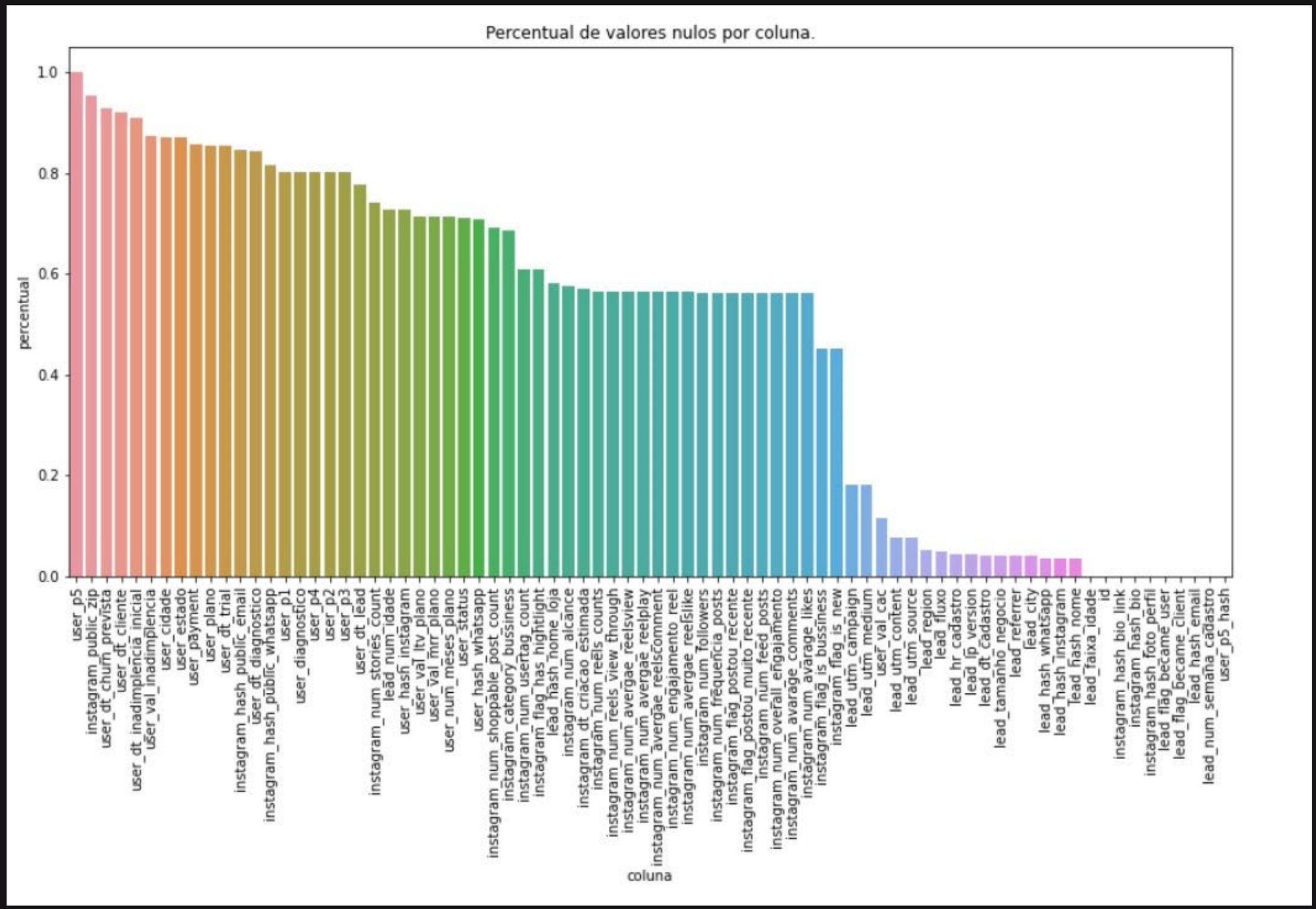
```
undefined      771
               219
eiro           122
                74
                57
              ...
                 1
                 1
                 1
                 1
a                1
uz               1
_city, Length: 705, dtype: int64
```

```
user_dt_lead
                      116
02/08/2022 18:40        1
02/11/2022 10:19        1
03/09/2022 10:47        1
07:05:26 06/03/2022     2
                      ...
31/07/2021 21:35        1
4/1/2022 18:40:44       2
4/5/2022 8:30:00        1
4/6/2022 14:46:00       1
5/6/2022 14:43:28       1
Name: id, Length: 695, dtype: int64
```

...

	column_name	total_duplicados	duplicados_not_na	duplicados_not_na_and_blank	duplicados_not_blank
	ash_nome_loja	1887	10	10	1887
	ad_hash_nome	162	54	54	162
	hash_instagram	122	14	14	122
	ad_hash_email	41	41	41	41
	hash_whatsapp	149	41	41	149
	_hash_bio_link	2452	2452	2452	2452
	ublic_whatsapp	2636	5	5	2636
	sh_public_email	2733	2	2	2733
8	user_hash_instagram	2366	18	18	2366
9	user_hash_whatsapp	2311	22	22	2311

Data Preparation



Análise de valores nulos por coluna.

Data Preparation

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3231 entries, 0 to 3230
Data columns (total 76 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              3231 non-null   int64
 1   lead_dt_cadastro                3100 non-null   datetime64[ns]
 2   lead_hr_cadastro                3088 non-null   object
 3   lead_num_semana_cadastro        3231 non-null   int64
 4   lead_hash_nome_loja             1353 non-null   object
 5   lead_hash_nome                  3122 non-null   object
 6   lead_hash_instagram             3122 non-null   object
 7   lead_hash_email                 3231 non-null   object
 8   lead_hash_whatsapp              3122 non-null   object
 9   lead_fluxo                      3071 non-null   category
 10  lead_tamanho_negocio            3100 non-null   category
 11  lead_utm_source                 2984 non-null   category
 12  lead_utm_medium                 2643 non-null   category
 13  lead_utm_campaign               2643 non-null   category
 14  lead_utm_content                2982 non-null   category
 15  lead_lp_version                 3089 non-null   category
 16  lead_referrer                   3100 non-null   category
 17  lead_city                       3100 non-null   category
 18  lead_region                     3066 non-null   category
 19  lead_flag_became_client         3231 non-null   int64
 20  lead_flag_became_user           3231 non-null   int64
 21  instagram_num_followers         1417 non-null   float64
```

Amostra dos tipos de dados.

Data Preparation

Houve a necessidade de:

- Desmembrar colunas;
- Normalização dos dados;
- Reagrupamento das colunas categóricas;
- Redução de categorias.

Modeling

1. Escolha das features para o modelo:

Features escolhidas:

- Melhor qualidade de dados
- Menor percentual de nulos
- Representam características descritivas dos Leads

Selecionadas 23 features do total de 76.

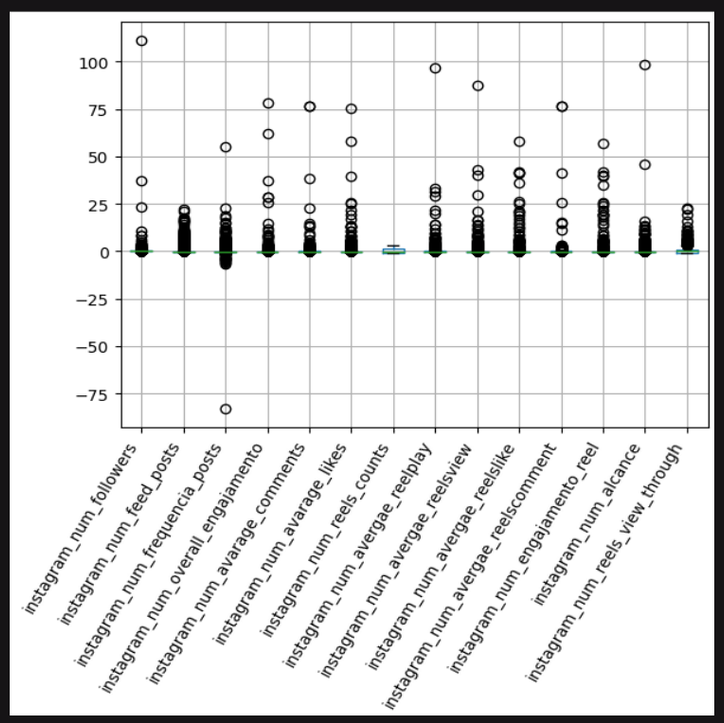
lead_hr_cadastro_ajustado	instagram_num_overall_engajamento
lead_fluxo_ajustado	instagram_num_avarage_comments
lead_tamanho_negocio_ajustado	instagram_num_avarage_likes
lead_utm_medium_ajustado	instagram_num_reels_counts
lead_lp_version_ajustado	instagram_num_avergae_reelplay
lead_referrer_ajustado	instagram_num_avergae_reelsview
lead_region_ajustado	instagram_num_avergae_reelslike
lead_num_trimestre_ajustado	instagram_num_avergae_reelscomment
lead_flag_became_client	instagram_num_engajamento_reel
instagram_num_followers	instagram_num_alcance
instagram_num_feed_posts	instagram_num_reels_view_through
instagram_num_frequencia_posts	

Modeling

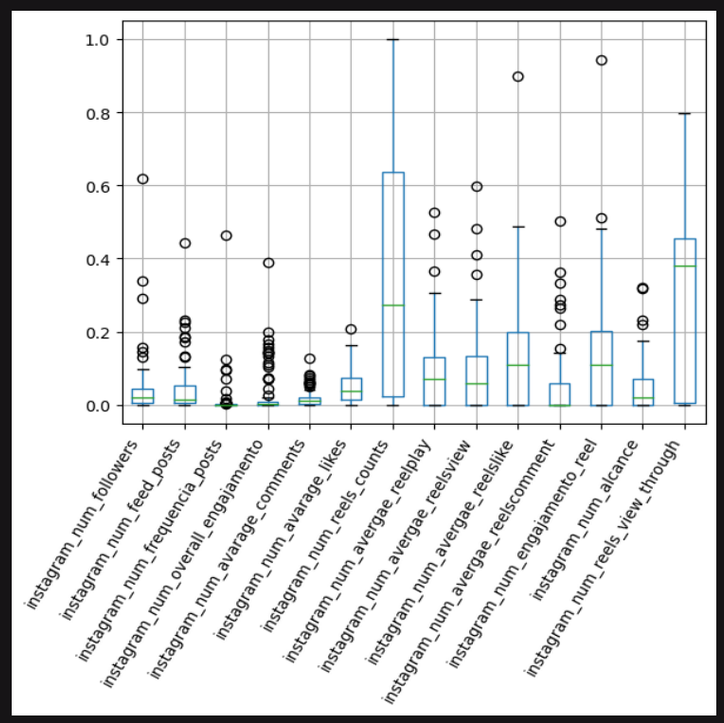
2. Feature Scaling dos dados numéricos:

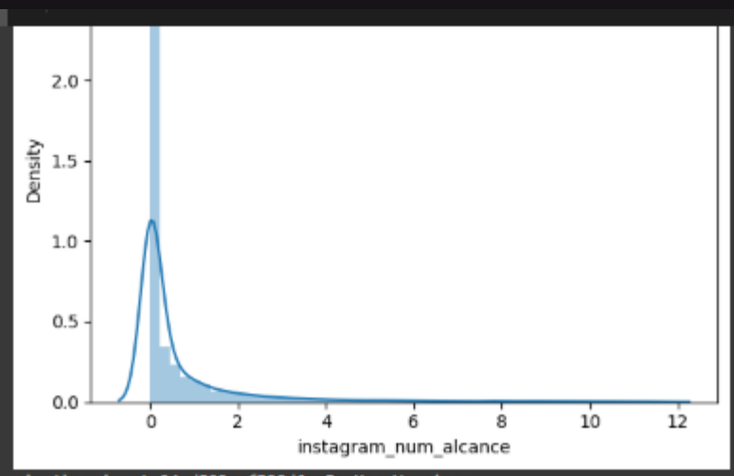
- Eliminação de dados nulos
- Tratamento de Outliers
- Método: MinMaxScaler
- Distribuição não Gaussiana

Antes



Depois





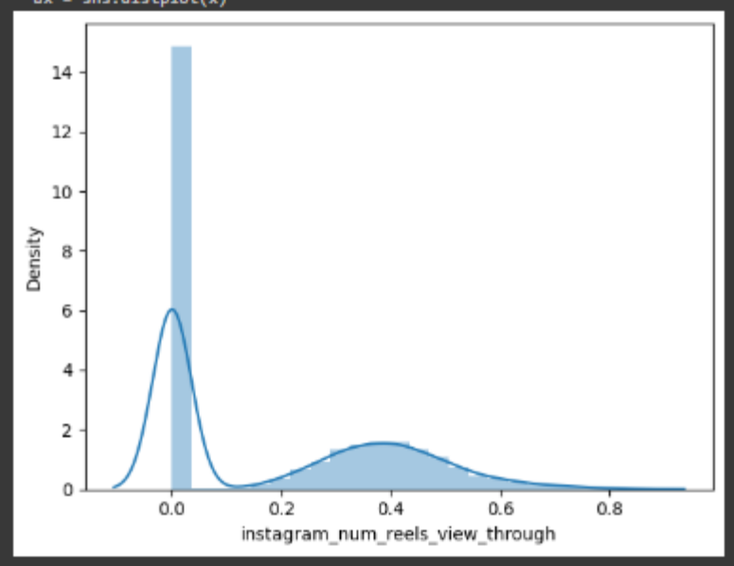

```
<ipython-input-24-d523ccf538d6>:5: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  ax = sns.distplot(x)
```



Modeling

2. Feature Scaling dos dados numéricos:

- Eliminação de dados nulos
- Tratamento de Outliers
- Método: MinMaxScaler
- Distribuição não Gaussiana

```python
# Normalizing numeric continuous variables

#Importing library
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

prepi_scoring.reset_index()
prepi_scoring[continuos_variables] = scaler.fit_transform(prepi_scoring[continuos_variables])
```

Modeling

3. **Get_Dummies em dados categóricos:**

- **Pré-processamento de dados categóricos**
- **Preenchimento de valores nulos**
- **Agrupamento de categorias**
- **Aplicação de método get_dummies**
- **Dataset resultante:** 71 colunas e 8503 linhas

```
[ ]  1 #Converting features to dummies
     2
     3 dummy_df = pd.get_dummies(modelling_dataset_client)
     4 dummy_df
```

Modeling

4. Rebalanceamento dos dados:

- Métodos testados: SMOTE, ADASYN, RO, RU, NM-1-2-3, CNN
- Melhores resultados: SMOTE e ADASYN
- Melhora de 7% no ROC AUC
- Proporção dos dados iniciais e finais:

```
y_train (No-ressample)

Positivos (1): 721 (10.60 %)
Negativos (0): 6081 (89.40 %)
```

```
y_train (Ressampled)

Positivos (1): 6087 (50.02 %)
Negativos (0): 6081 (49.98 %)
```

```
ROC_AUC Treino:  99.79195561719834
ROC_AUC Teste:  54.30120678408349
```

```
ROC AUC Treino:  99.98355533629338
ROC AUC Teste:  61.839530332681015
```

```
[298]  1 # SMOTE - Rebalancing dataset
       2
       3 sm = SMOTE(random_state=42)
       4 X_train_res, y_train_res = sm.fit_resample(X_train,y_train)
```

Modeling

5. Técnica de modelagem:

- **Modelos avaliados:** Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, MLP Classifier e Kneighbors Classifier
- **Melhor modelo:** Random Forest
- **Utilizamos 80% dos dados para treinamento e 20% de teste.**

```
[299]  # Training the model with SMOTE rebalanced columns
       model_rf = RandomForestClassifier(bootstrap = False,
                                         max_depth = None,
                                         max_features = 'sqrt',
                                         n_estimators = 100,
                                         min_samples_split = 4)
       model_rf.fit(X_train_res,y_train_res)
```

```
RandomForestClassifier
RandomForestClassifier(bootstrap=False, min_samples_split=4)
```

```
[300]  y_pred_train_res = model_rf.predict(X_train_res)
       y_pred_test= model_rf.predict(X_test)
```

Evaluation

1. Avaliação dos resultados:

- O modelo escolhido (Random Forest Classifier + SMOTE) apresentou resultados conforme abaixo, com bom desempenho em métricas como acurácia, acurácia balanceada, ROC AUC e F1 Score.

```
Acc Treino:  99.96711067258674
Acc Teste:  89.2416225749559
================================================
F1 Treino:  99.96711066902908
F1 Teste:  88.01685995842932
================================================
ROC AUC Treino:  99.96711067258674
ROC AUC Teste:  60.905903457273325
================================================
Confusion Matrix
 [[1475   58]
 [ 125   43]]
================================================
Classification Report
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      1533
           1       0.43      0.26      0.32       168

    accuracy                           0.89      1701
   macro avg       0.67      0.61      0.63      1701
weighted avg       0.87      0.89      0.88      1701

Precision: Percentage of correct positive predictions relative to total positive predictions
Recall: Percentage of correct positive predictions relative to total actual positives
F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model
```

Evaluation

2. Geração do Lead Scoring:

- Extração da probabilidade de virar cliente (1) pelo predict_proba
- Organização em ordem decrescente para geração ranking
- Rastreabilidade do Lead através do ‘id’

```
y_pred_train_res = model_rf.predict(X_train_res)
y_pred_test= model_rf.predict(X_test)
y_pred_proba_test = model_rf.predict_proba(X_test)
```

```
y_pred_proba_test = pd.DataFrame(y_pred_proba_test, columns=['y_pred_proba_0','y_pred_proba_1'])
y_pred_proba_test
```

```
X_test_final = X_test.copy()
X_test_final['y_pred_test'] = y_pred_test
X_test_final['y_test'] = y_test


X_test_final = pd.concat([X_test_final, y_pred_proba_test.set_index(X_test_final.index)], axis=1)
```

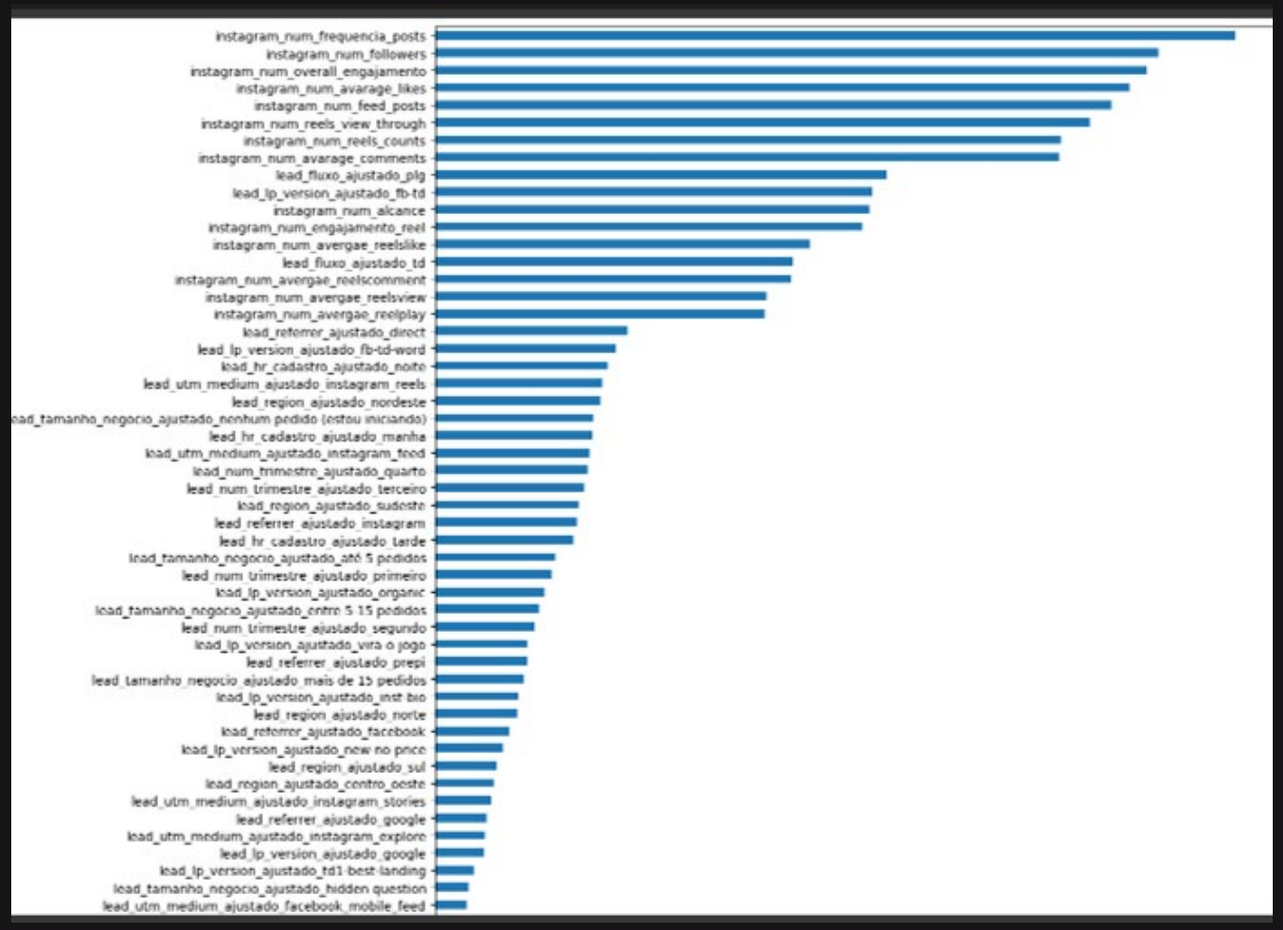
```
X_test_final = X_test_final.sort_values(by='y_pred_proba_1',ascending=False)
X_test_final.head(20)
```

id	y_pred_test	y_pred_proba_0	y_pred_proba_1
27888	1	0.00	1.00
30723	1	0.07	0.93
25847	1	0.14	0.86
10138	1	0.16	0.84
10987	1	0.18	0.82
9921	1	0.18	0.82
18226	1	0.19	0.81
8723	1	0.19	0.81
18423	1	0.20	0.80
11326	1	0.22	0.78

Evaluation

3. Escolha das features para o modelo:

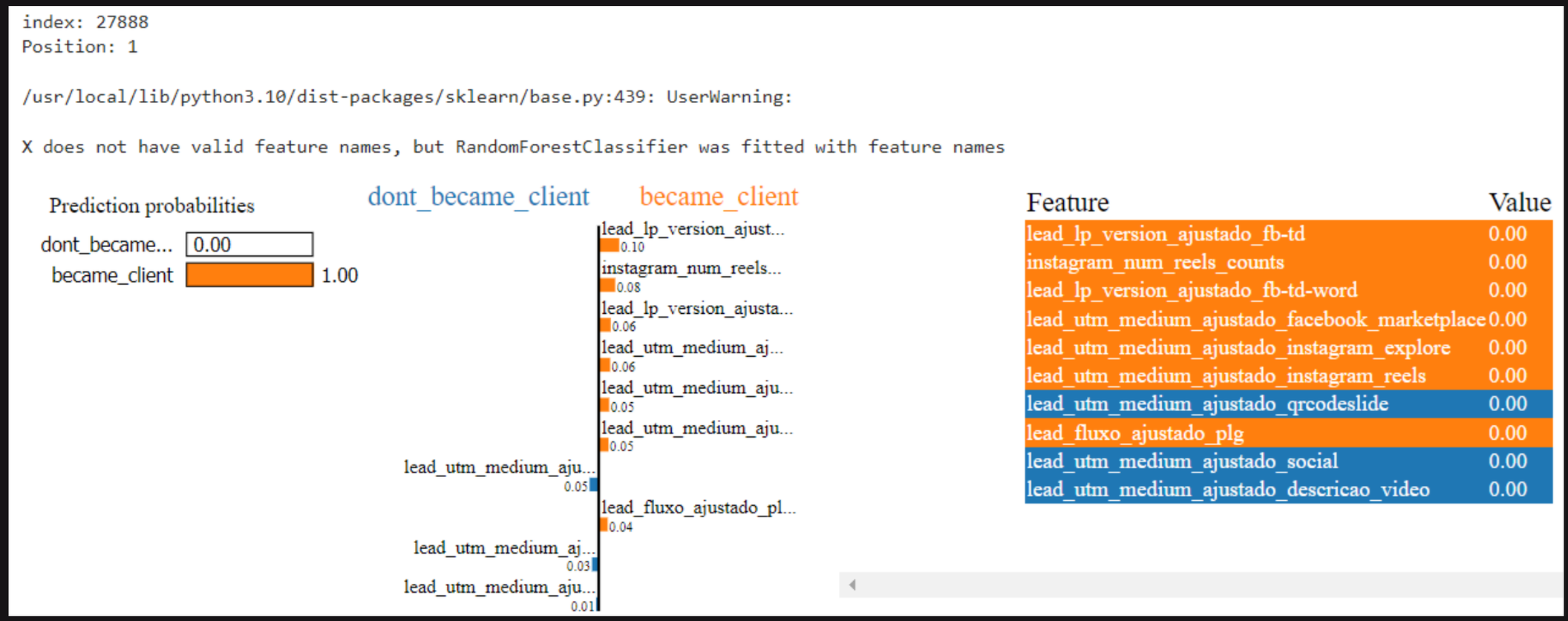
- Extração das importâncias das features para resposta do modelo.
- Quais ‘características’ dos Leads impactam para previsão



Evaluation

3. Escolha das features para o modelo:

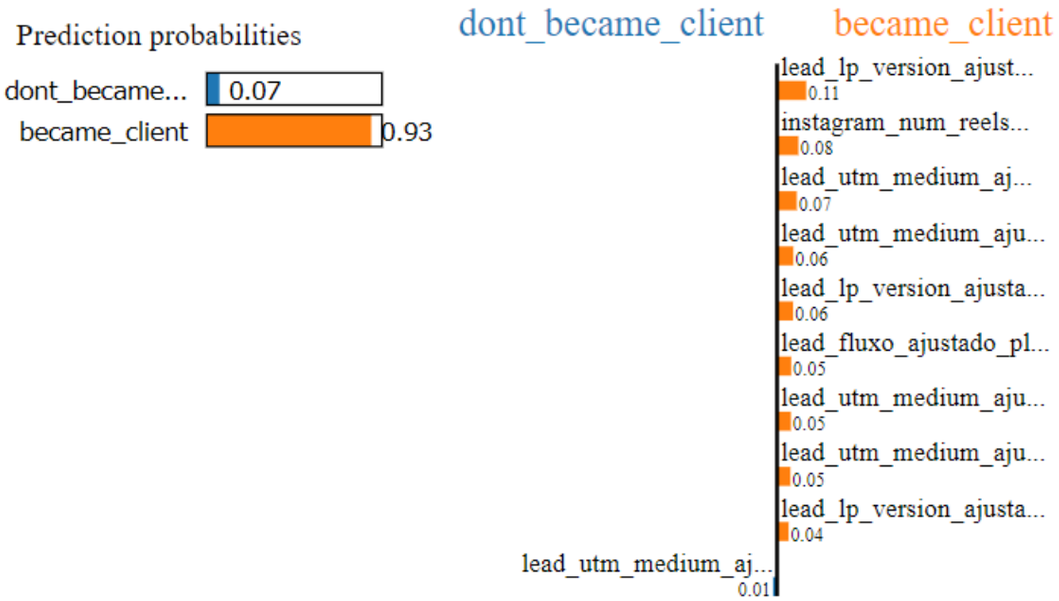
- Extração individual dos 10 primeiros do Lead Scoring



Análise utilizando biblioteca “Lime”.

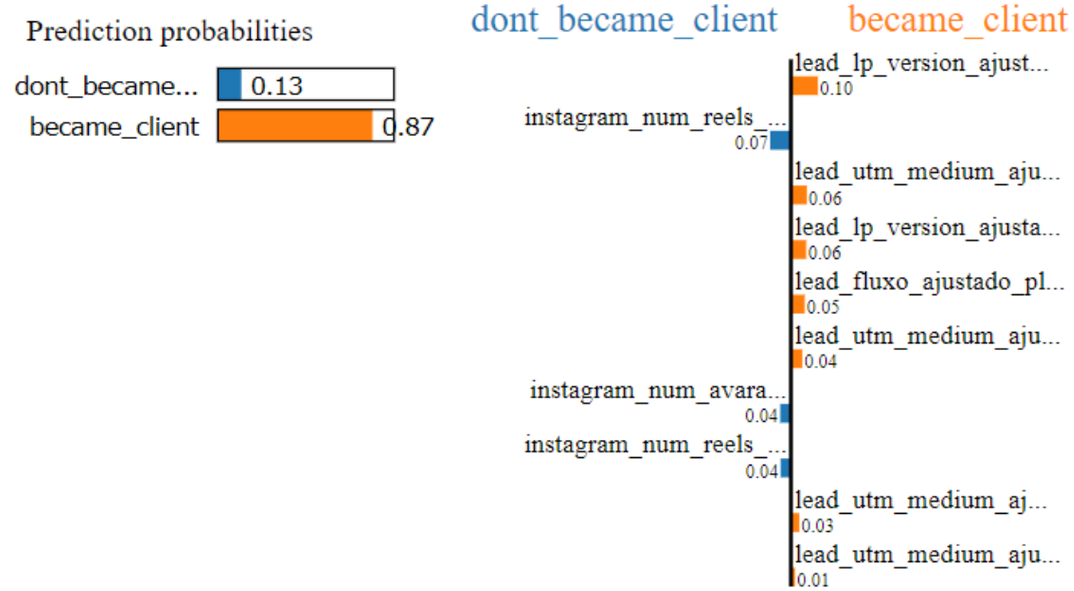
Evaluation

index: 30723
Position: 2



Feature	Value
lead_lp_version_ajustado_fb-td	0.00
instagram_num_reels_counts	0.00
lead_utm_medium_ajustado_facebook_marketplace	0.00
lead_utm_medium_ajustado_descricao_video	0.00
lead_lp_version_ajustado_fb-td-word	0.00
lead_fluxo_ajustado_plg	0.00
lead_utm_medium_ajustado_instagram_explore	0.00
lead_utm_medium_ajustado_instagram_reels	0.00
lead_lp_version_ajustado_new-no-price	0.00
lead_utm_medium_ajustado_social	0.00

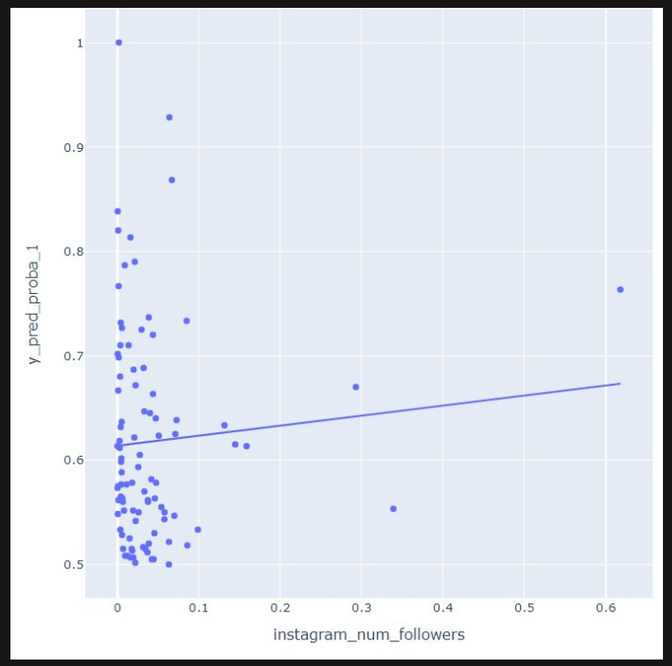
index: 18423
Position: 3

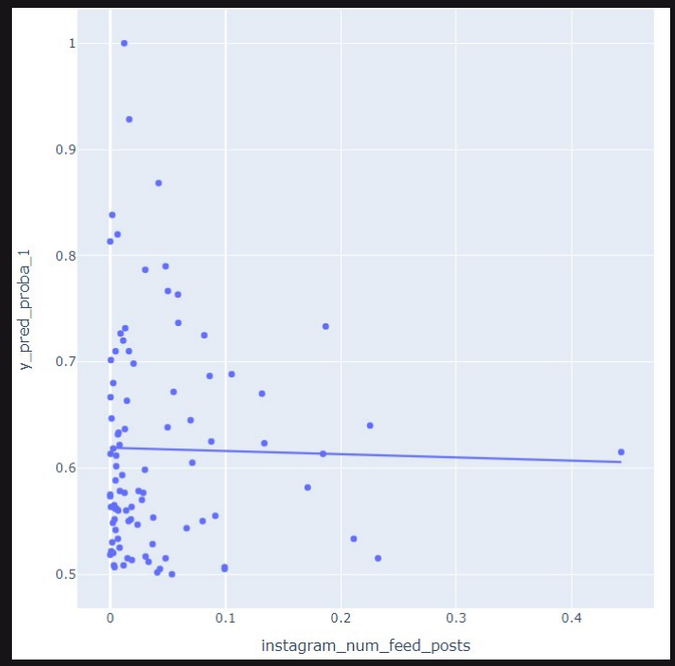


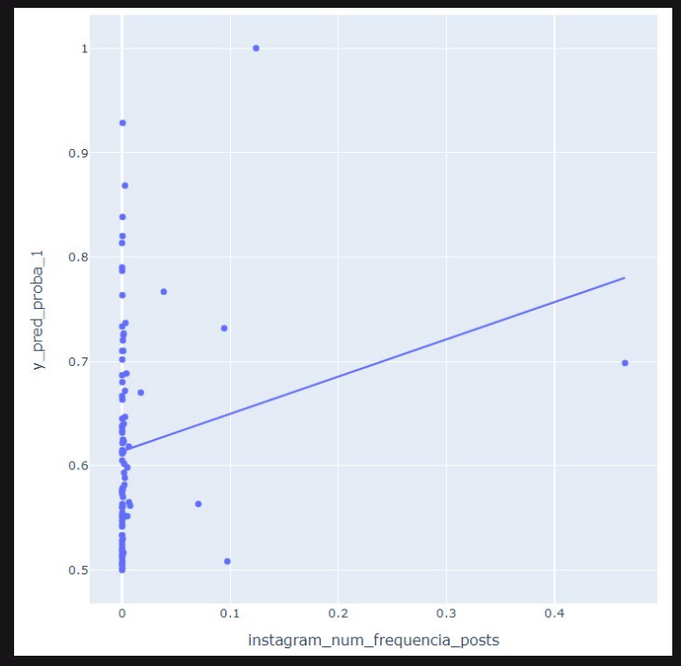
Feature	Value
lead_lp_version_ajustado_fb-td	0.00
instagram_num_reels_counts	0.55
lead_utm_medium_ajustado_instagram_reels	0.00
lead_lp_version_ajustado_fb-td-word	0.00
lead_fluxo_ajustado_plg	0.00
lead_utm_medium_ajustado_descricao_video	0.00
instagram_num_avarage_likes	0.10
instagram_num_reels_view_through	0.52
lead_utm_medium_ajustado_whatsapp	0.00
lead_utm_medium_ajustado_facebook_desktop_feed	0.00

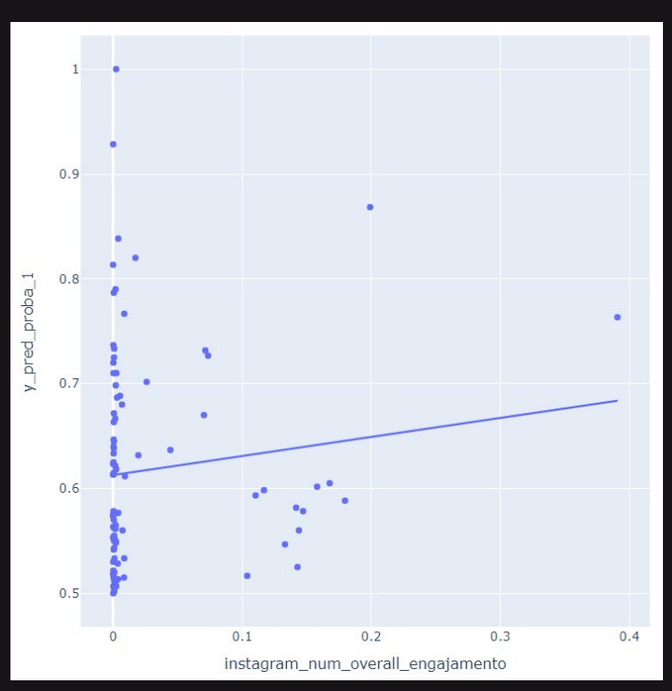
Análise utilizando biblioteca “Lime”.

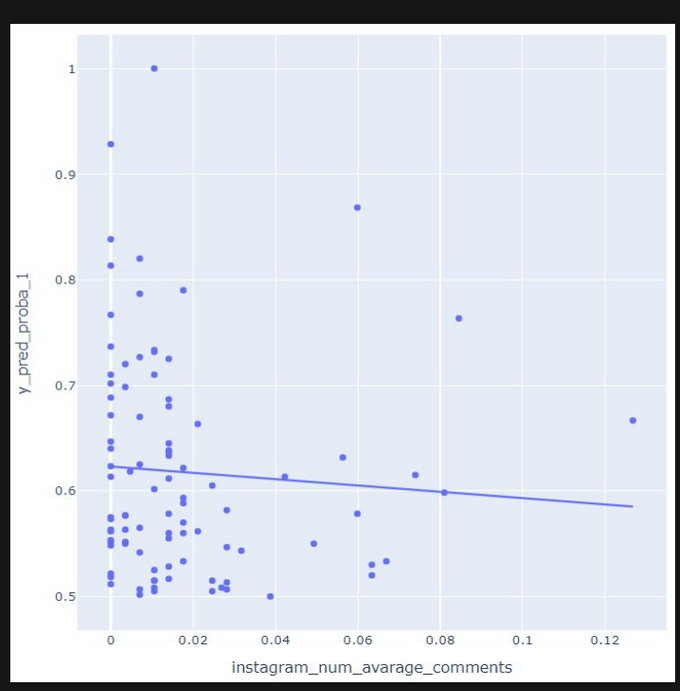
Evaluation

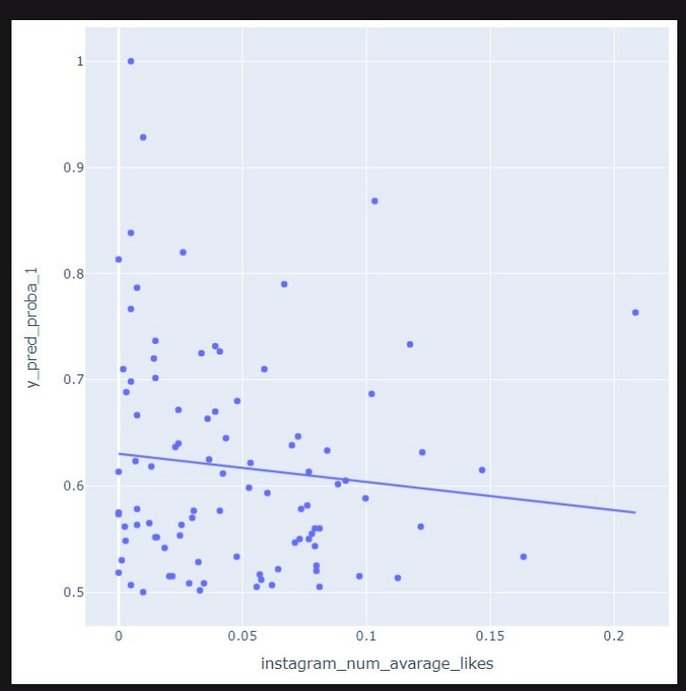






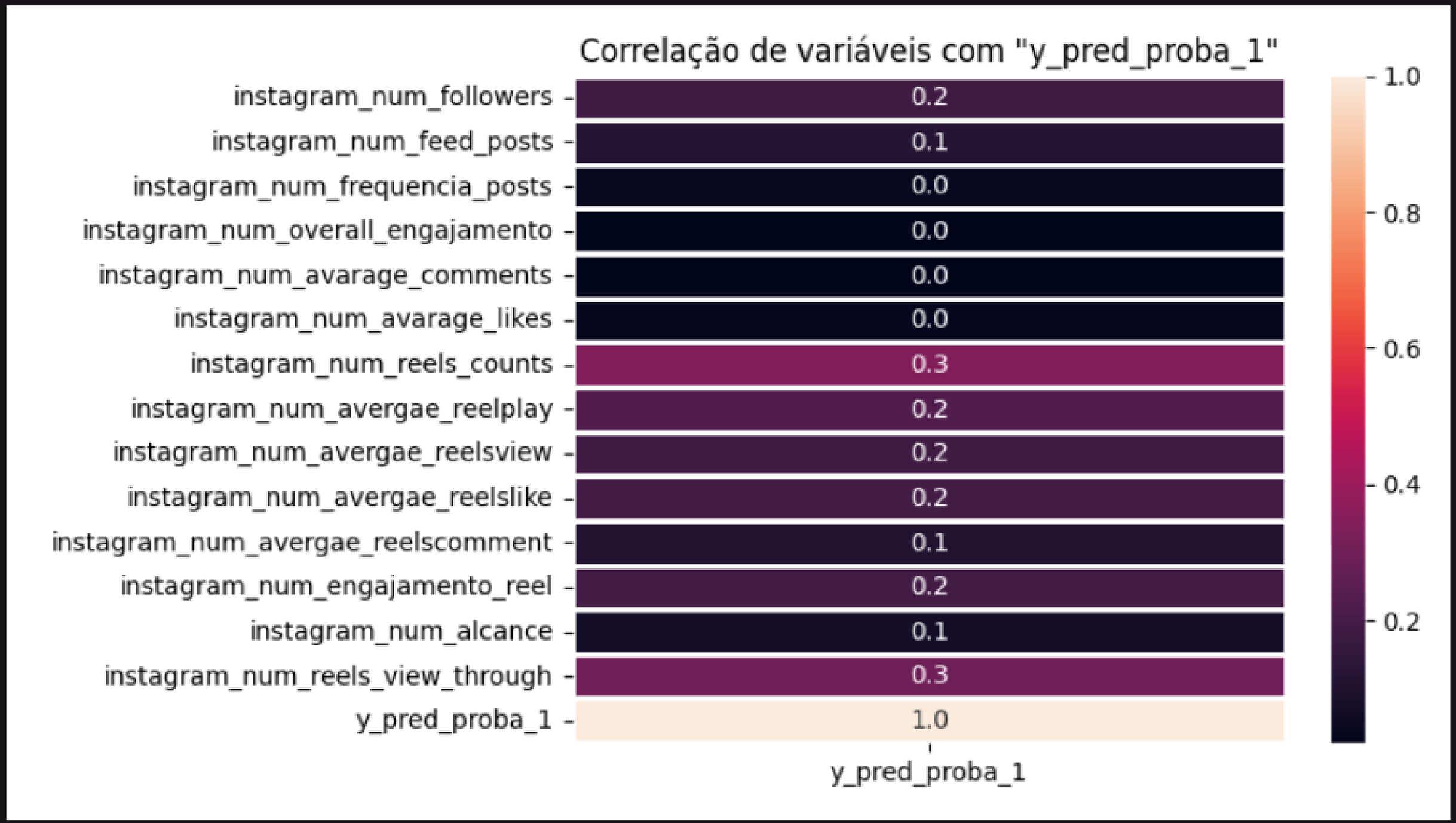






Análise de distribuição de valores de features numéricas normalizadas versus a probabilidade do lead virar cliente.

Evaluation



Análise de correlação de variáveis com a probabilidade do lead virar cliente.

Evaluation

4. Revisão do processo:

- Todas as etapas do projeto foram concluídas conforme o planejado, exceto a etapa de Deployment;
- O modelo desenvolvido alcançou resultados satisfatórios para previsões negativas, porém com oportunidade de evolução nas positivas;
- Os resultados oferecem apoio na tomada de decisões da PREPI;

Deployment

Sugerimos a aplicação de um *score* (nota) do *lead* (cliente), que utilize os dados coletados dos *leads* e considere o resultado previsto das métricas mais importantes para o negócio, como:

- Probabilidade do lead virar cliente pagante
- CAC previsto
- LTV anual previsto
- Probabilidade de churn anual

Não foram realizados implementações de deploy do modelo.

Deployment

Utilizando o método de **BSC** (Balance Score Card) é possível calcular um índice de 0 a 100%, para cada *lead*, ou grupo de *leads*, definindo assim um *lead score* geral que englobe várias premissas de sucesso na captação e seleção de clientes da empresa.

Exemplo:

Lead Score = (Nota1 * Peso1) + (Nota2 * Peso2) + (NotaN * PesoN) ...

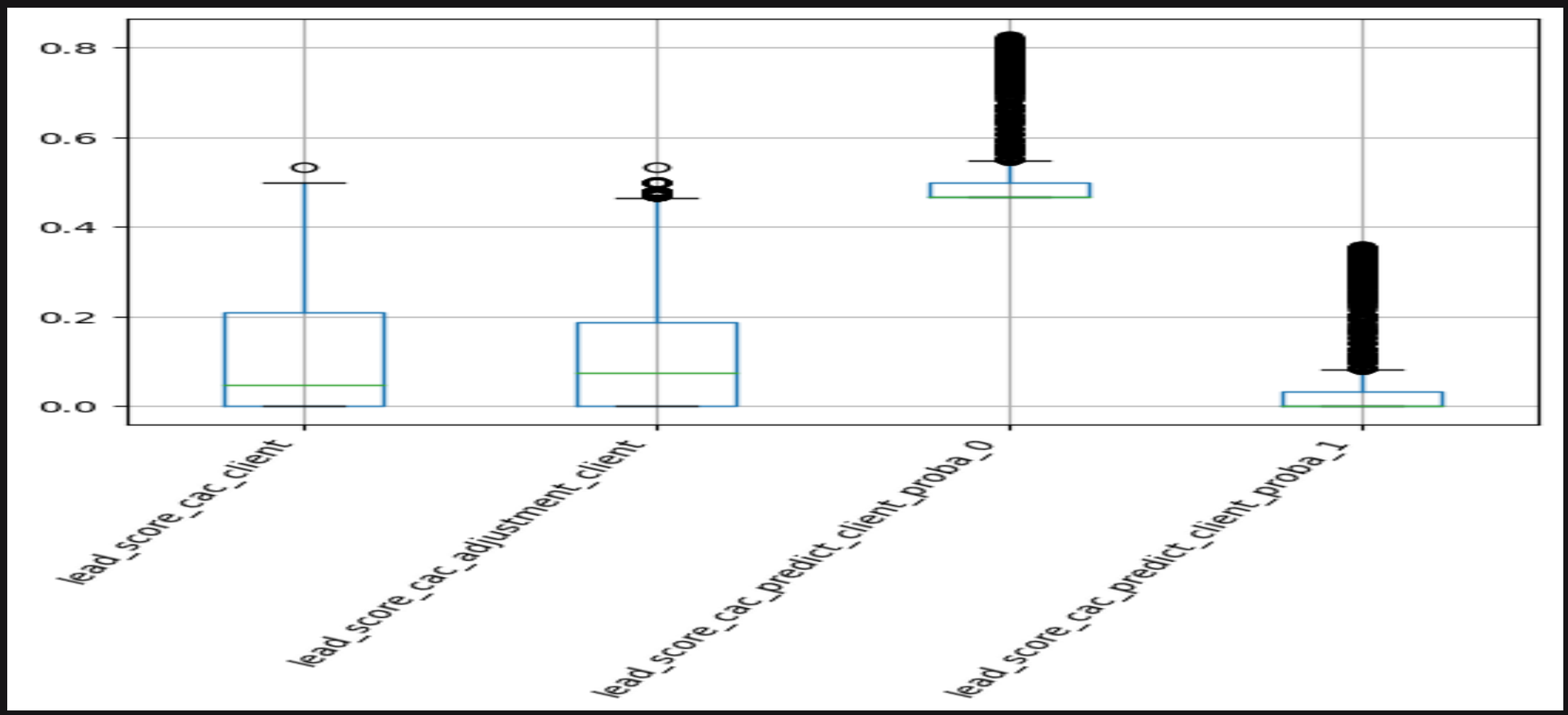
p/ **NotaN** = PrevisãoN / TargetN .. ou .. **NotaN** = % PrevisãoN

p/ 0 < **NotaN** > 1

p/ 1 = **Peso1 + Peso2 + PesoN** + ...

Deployment

	count	mean	std	min	25%	50%	75%	max
lead_score_cac_client	30707.0	0.120182	0.138881	0.000000e+00	0.000000e+00	0.048333	0.209667	0.533333
lead_score_cac_adjustment_client	34115.0	0.115474	0.132597	0.000000e+00	0.000000e+00	0.073667	0.186333	0.533333
lead_score_cac_predict_client_proba_0	34115.0	0.528800	0.111835	4.666667e-01	4.666667e-01	0.466669	0.500000	0.824234
lead_score_cac_predict_client_proba_1	34115.0	0.062133	0.111835	2.156789e-18	8.483117e-09	0.000002	0.033333	0.357567



Exemplos de aplicação do BSC.

Considerações Finais

Implantação de etapas de tratamento de dados individuais para cada variável.

Definição de negócio para preenchimento de dados faltantes (nulos) para cada variável.

Criação de um modelo otimizado para cada métrica de negócio.

Definição da importância de cada métrica de negócio e seu respectivo peso.

Definição de um “base line” para avaliação do sucesso de implementação do Lead Score (back test, taxa de conversão, CAC médio, etc...).

Necessário realizar revisão das features utilizadas na tentativa de otimizar as previsões positivas do modelo.

Obrigado