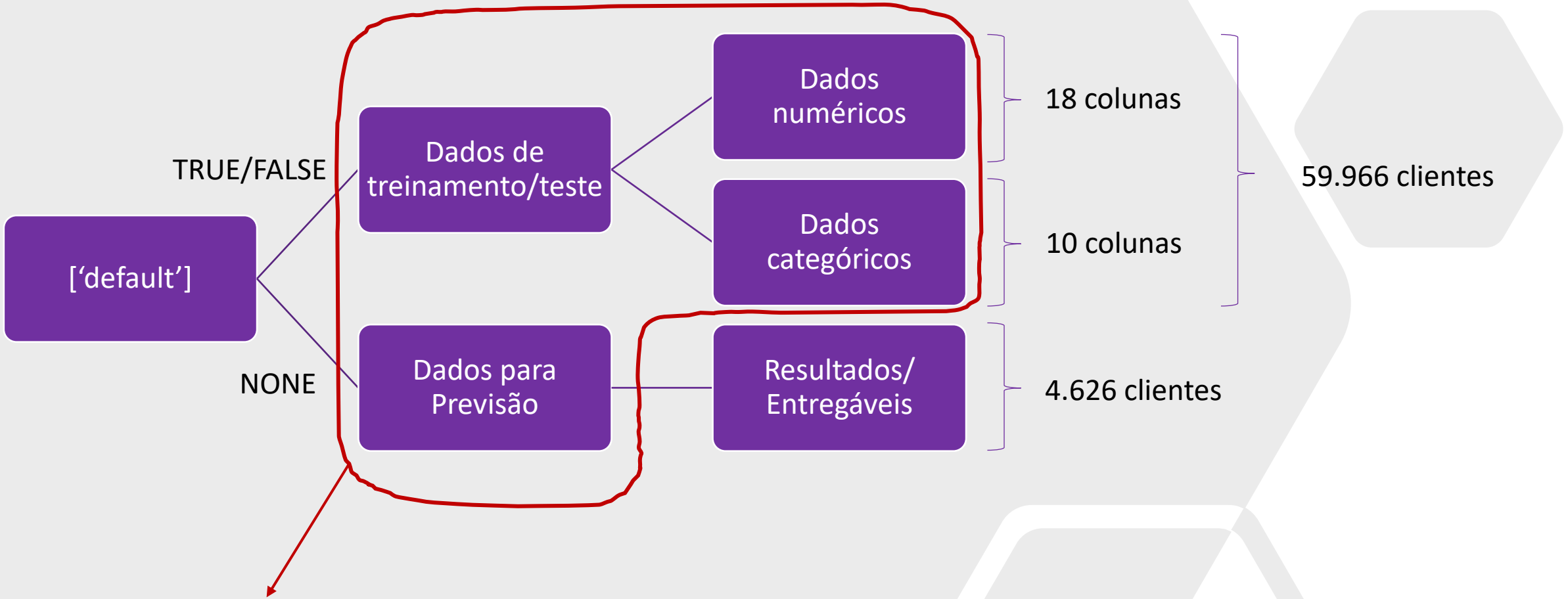
Estudo de Caso: Risco de Crédito "Grande Banco Brasileiro"

Processo de Seleção Accenture - Fase 3

Preparado por: Armando Alvarez Rolins

Resumo dos Dados



Feature Engineering → Modelo → Resultado

Feature Engineering

***Outliers* Categóricos:** Categorias que aparecem com pouca frequência agrupadas como '*other*' (*Capping*) e *features* com pouca importância de acordo com o modelo.

['score_1'] < 10.000 | ['score_2'] < 1.000 | ['reason'] < 1.000 | ['state'] < 500 | ['zip'] < 500 | ['channel'] | ['job_name'] < 100 | ['real_state'] < 1.000 | ['facebook_profile'] < 1.000 | ['sign'] < 1.000

***Creating*:** *Features* que parecem relevantes para o problema (subjetivo): $d2i = \frac{amount\ borrowed}{income}$ $l2i = \frac{credit\ limit}{income}$

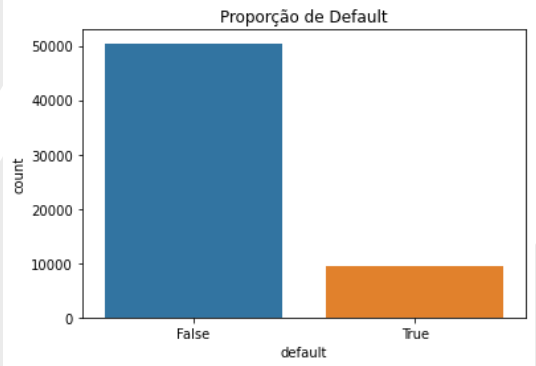
***Imputing*:** Usando valores médios para *features* numéricas.

```
In [161]: print(X_num_train)
       score_3     score_4  ...  d2i_ratio  i2i_ratio
0        710.0  104.174961  ...   0.334468   1.042041
1        330.0   97.880798  ...   0.218325        NaN
2        360.0   97.908925  ...   0.442021        NaN
3        120.0  100.434557  ...   0.500582        NaN
4        330.0  103.774638  ...   0.266881   0.359518
...        ...         ...  ...        ...        ...
64587    440.0   99.087197  ...   0.126828        NaN
64588    230.0   96.473000  ...   0.128418   0.203570
64589    320.0   93.221044  ...   0.350123   0.186527
64590    590.0   96.613431  ...   0.182691        NaN
64591    190.0   98.331847  ...   0.397998   0.290172

[59966 rows x 18 columns]
```

```
In [162]: print(X_num_train_imp)
       score_3     score_4  ...  d2i_ratio  i2i_ratio
0        710.0  104.174961  ...   0.334468   1.042041
1        330.0   97.880798  ...   0.218325   0.494824
2        360.0   97.908925  ...   0.442021   0.494824
3        120.0  100.434557  ...   0.500582   0.494824
4        330.0  103.774638  ...   0.266881   0.359518
...        ...         ...  ...        ...        ...
59961    440.0   99.087197  ...   0.126828   0.494824
59962    230.0   96.473000  ...   0.128418   0.203570
59963    320.0   93.221044  ...   0.350123   0.186527
59964    590.0   96.613431  ...   0.182691   0.494824
59965    190.0   98.331847  ...   0.397998   0.290172

[59966 rows x 18 columns]
```

***Encoding*:** *One-Hot Encoding* para *features* categóricas.

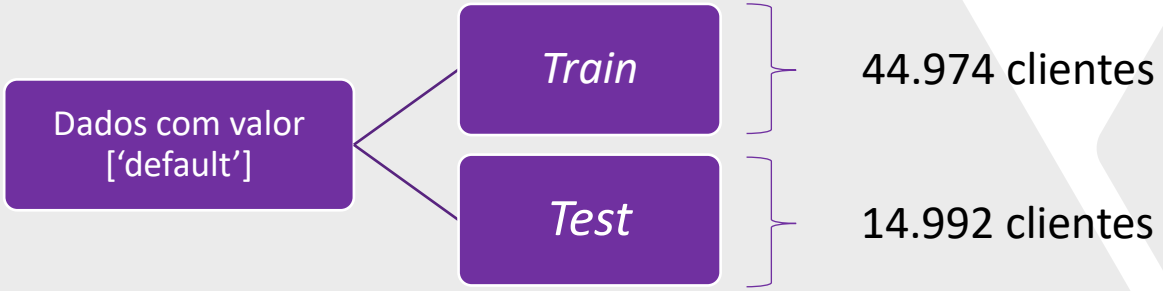
***Random undersampling*:** Aplicação de *random undersampling* (pela carga computacional menor em comparação ao *oversampling*) para equilibrar as classes de default.



Resultados

Split*: *Test-Train Split

- **random_state:** 0
- **test_size:** 0.25 (padrão)
- **train_size:** 0.75 (padrão)



Usando GridSearchCV para encontrar os melhores parâmetros:

Regressão Logística:

- **Penalty**: L1 Lasso Regression
- **C:** 0,6158 (Regularização)
- **Solver:** Liblinear

Recall: 64,1%

	prob_de_default
count	4626.000000
mean	0.655098
std	0.105218
min	0.175884
25%	0.588668
50%	0.665948
75%	0.745189
max	0.849011

***Random Forest*:**

- **max_depth**: 14
- **max_features:** 12
- **min_sample_leaf:** 8
- **n_estimators:** 150

Recall: 64,9%

	prob_de_default
count	4626.000000
mean	0.634899
std	0.098304
min	0.246543
25%	0.573137
50%	0.654418
75%	0.707388
max	0.877282

Refinamentos e Melhorias

Pontos para refinar os resultados:

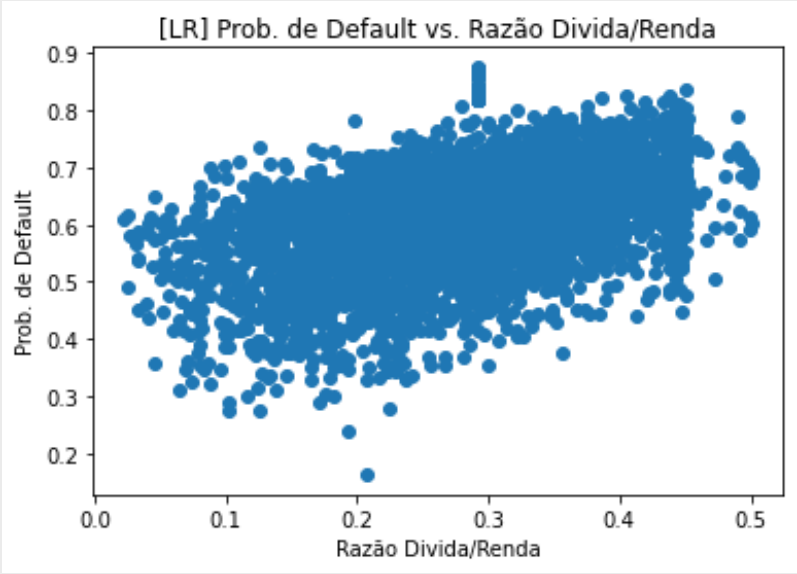
- Selecionar mais cuidadosamente (com análise técnica) quais *features* remover e fazer um refinamento do '*capping*'.
- Implementar '*Bucketing*' para *features* numéricas com alta variação de ordens de grandeza.
- Fazer tratamento de dados não preenchidos.
- Condensar *features* que já estão incluídas em outras (ex. razão dívida/renda).
- Fazer uma varredura mais fina (com mais tempo e/ou poder de processamento) de parâmetros iniciais dos modelos.

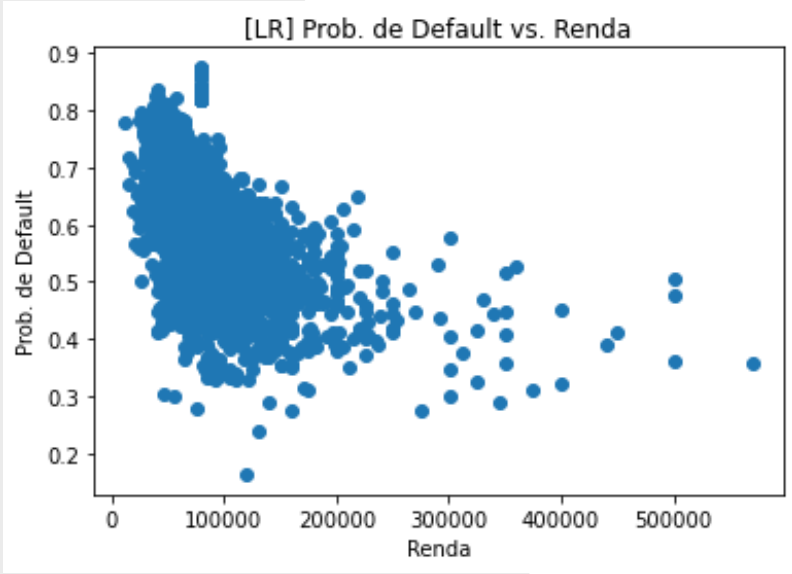
O resultado final encontra-se nos arquivos:

'../prob_de_default_random_forest.csv' e '../prob_de_default_logistic_regression.csv'

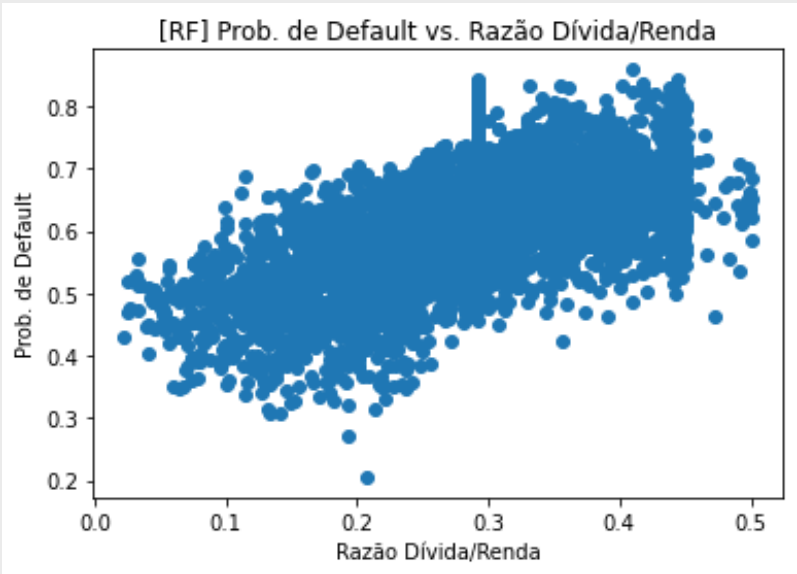
Anexo A

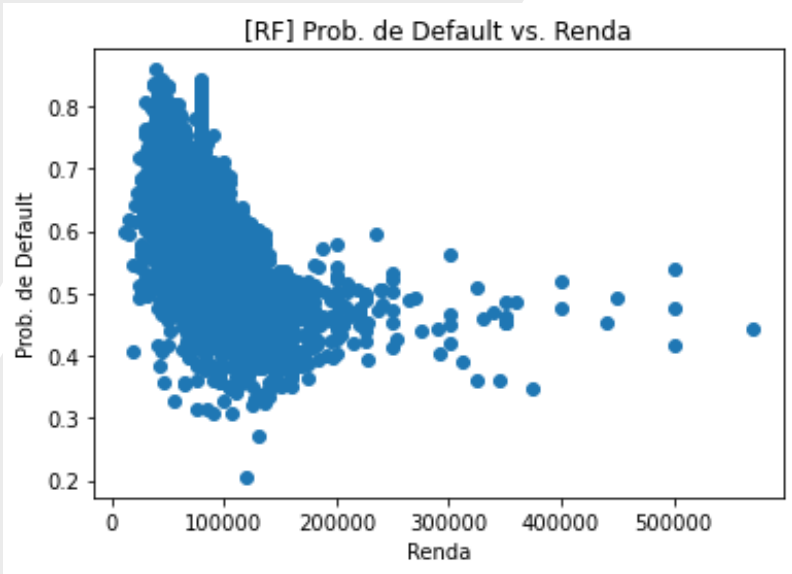
Regressão Logística





Random Forest





Anexo B

Importância de cada
feature pelo *Random Forest*

feature	importance
d2i_ratio	0.094323
income	0.089690
borrowed_in_months	0.078008
amount_borrowed	0.066929
score_6	0.062908
score_5	0.062269
score_4	0.062260