Projeto Credit Approval Data Set

Marcus Vinicius Bernardo 12/2022

Sumário

1.Introdução

O projeto consiste em uma análise de dados e a criação de um modelo de Machine learning com a principal finalidade de aprovar concessão de crédito. A base de dados está disponível na UCI, Credit Approval Data Set. Os dados foram ocultos para proteger a privacidade dos mesmos. Nesse documento terá uma abordagem nas principais modelagem e aplicação realizada neste *dataset* a fim de obter um bom resultado.

2. Sobre os dados

A base de dados contém 15 atributos, sendo:

A1	b, a
A2	continuous
A3	continuous
A4	u, y, l, t.
A5	g, p, gg
A6	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
A7	v, h, bb, j, n, z, dd, ff, o.
A8	continuous
A9	t, f.
A10	t, f.
A11	continuous
A12	t, f.

A13	g, p, s.
A14	continuous
A15	continuous
A16	+,-

Como dito na introdução, os dados foram alterados para aumentar a privacidade. Assim, os dados foram baixados dessa maneira.

	b	30.83	0	u	g	w	v	1.25	t	t.1	01	f	g.1	00202	0.1	+
0	a	58.67	4.460	u	g	q	h	3.04	t	t	6	f	g	00043	560	+
1	a	24.50	0.500	u	g	q	h	1.50	t	f	0	f	g	00280	824	+
2	b	27.83	1.540	u	g	w	v	3.75	t	t	5	t	g	00100	3	+
3	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	00120	0	+
4	b	32.08	4.000	u	g	m	v	2.50	t	f	0	t	g	00360	0	+

Figura 1: Base de dados do projeto

3. Exploração dos dados

Nessa etapa, o principal objetivo era conhecer mais sobre os dados. Foi analisada informações dos dados, no qual confirmava os tipos de dados que continham nesse dataset. Ademais, foi explorado valores nulos, estatísticas como correlação e variância.

	0	1.25	01	0.1
0	1.000000	0.298714	0.271003	0.122935
1.25	0.298714	1.000000	0.322247	0.051267
01	0.271003	0.322247	1.000000	0.063616
0.1	0.122935	0.051267	0.063616	1.000000

Figura 2: Tabela de correlação dos dados

Nessa etapa, ficou evidente que uma (1) *feature* tinha uma alta variância. O procedimento para tratar essa variância irá estar no próximo tópico. Abaixo a variância dos campos numéricos.

```
0       2.478517e+01
1.25    1.121405e+01
01      2.367971e+01
0.1     2.718312e+07
dtype: float64
```

Figura 3: Tabela de variância

4. Visualização

Nessa fase, o foco era nos plots de gráfico. Busquei analisar padrões de distribuição, análise de *outliers* através de boxplots e percentis.

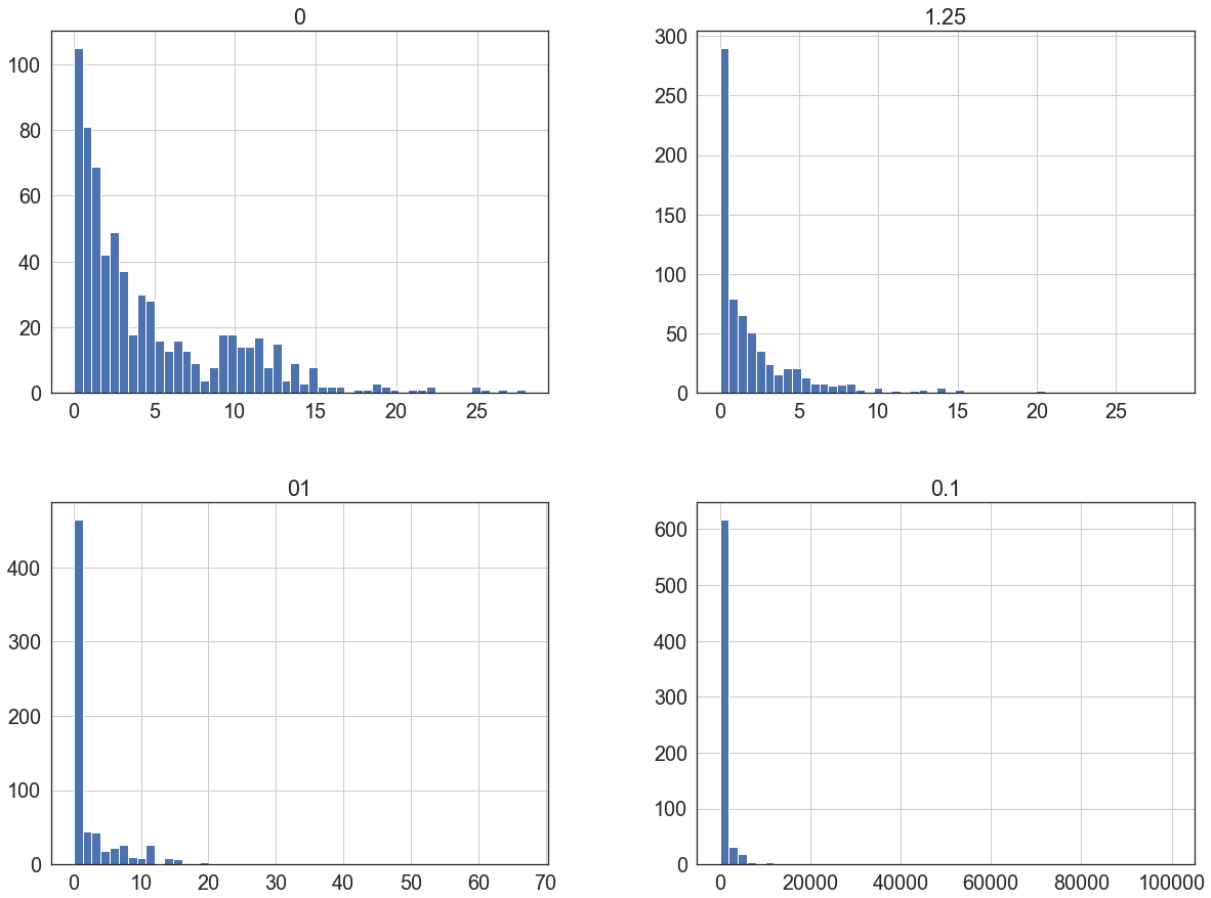

*Figura 5: Análise da distribuição dos dados numéricos através do **Histograma**.*

Nota-se que os dados não seguem uma distribuição normal. Além disso, nos outliers a coluna dita acima que contém uma alta variância também contém outliers em maior quantidade e escala. Abaixo, outlier do campo 0.1. Foi plotado separadamente a fim de facilitar a visualização.

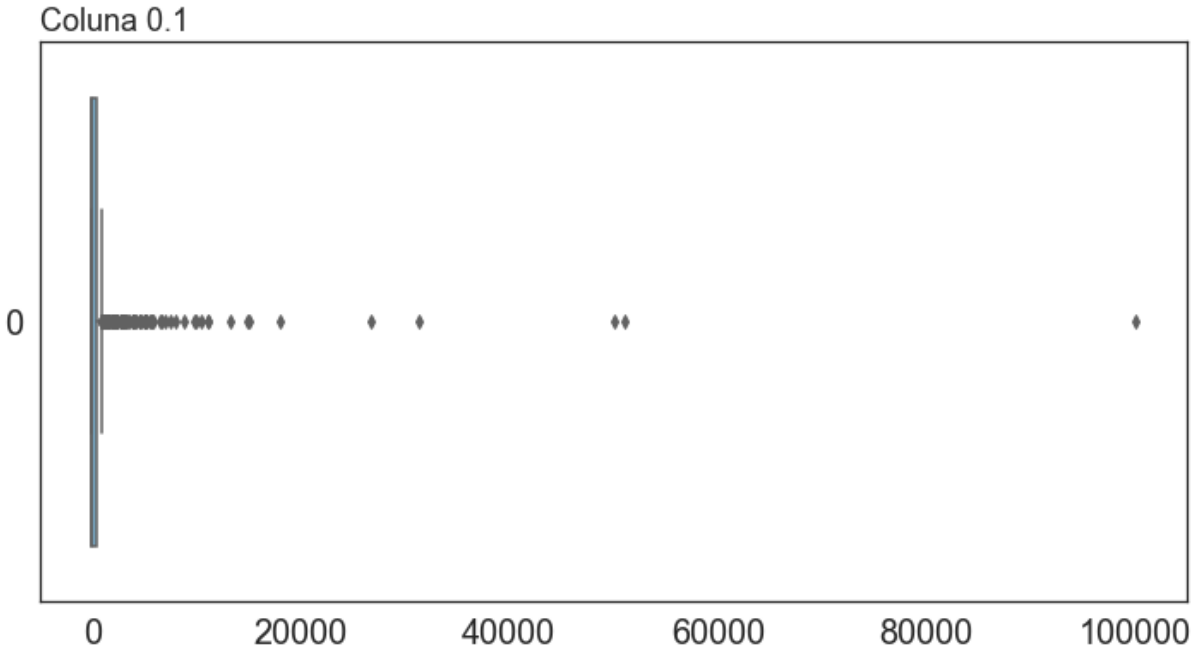

Figura 7: Gráfico de caixas da coluna 0.1 - Outliers

5. Pré Processamento

Nessa etapa o objetivo era realizar todos os ajustes nos dados a fim de projetar um bom modelo de Machine Learning. Na coluna com alta variância, 0.1, foi realizada uma transformação logarítmica para conter a variância. Analisando o novo boxplot após a transformação percebe-se uma estabilização da coluna, conforme a figura abaixo.

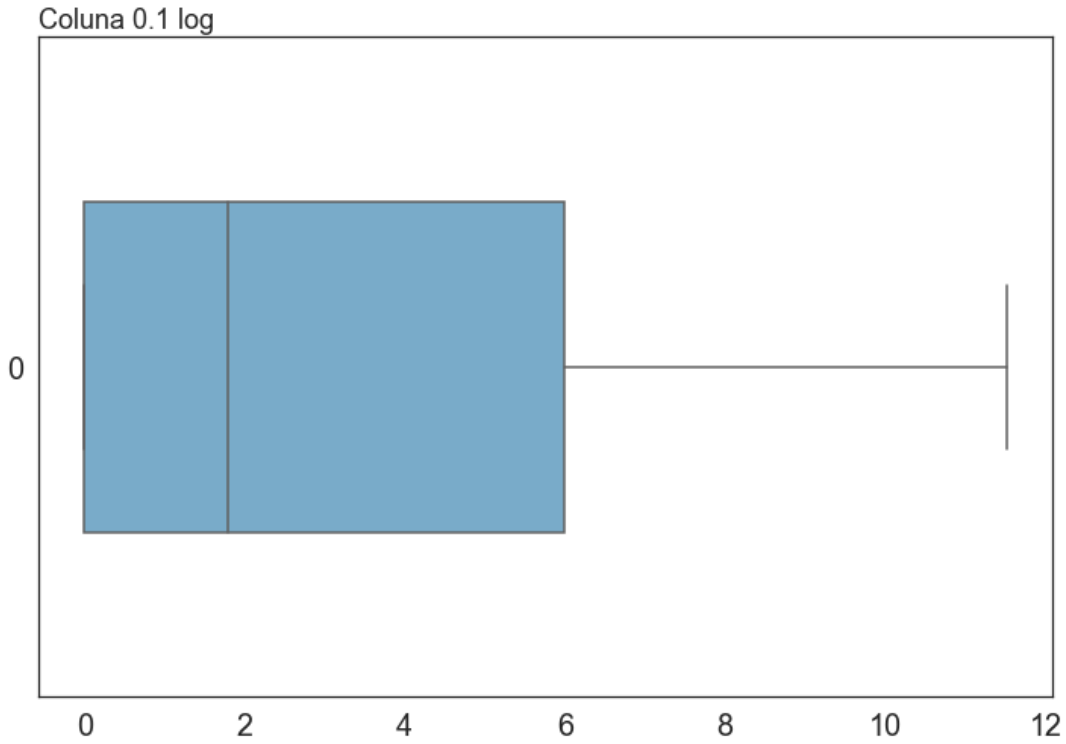

Figura 8: Boxplot da coluna transformada

Outrossim, foi analisada a distribuição dos dados da variável *target*. Como analisado, foi decidido não balancear os dados dessa coluna, pois seguia uma distribuição relativamente boa. E também foi criado uma pipeline aplicando **One Hot Encoder** nas colunas categóricas, e **Standard Scaler** para padronizar os dados numéricos.

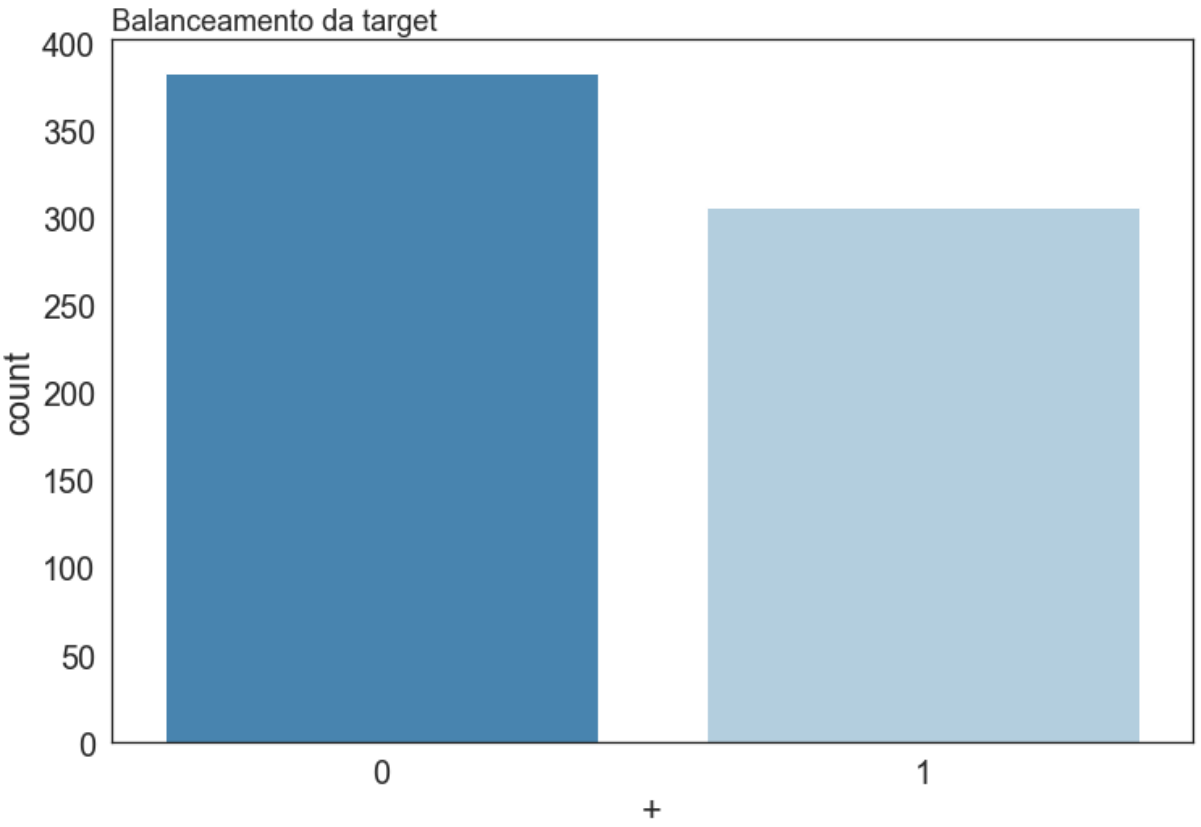

Figura 9: Gráfico da distribuição dos dados alvos

6. Modelo

Os modelos escolhidos foram Random Forest Classifier e Logistic Regression. A avaliação do modelo consistiu em *acuracy*, *confusion matrix*, *classifcation report* e *Cross Validate*. O modelo Random Forest teve uma acurácia maior 87% contra 83% do Logistic Regression.

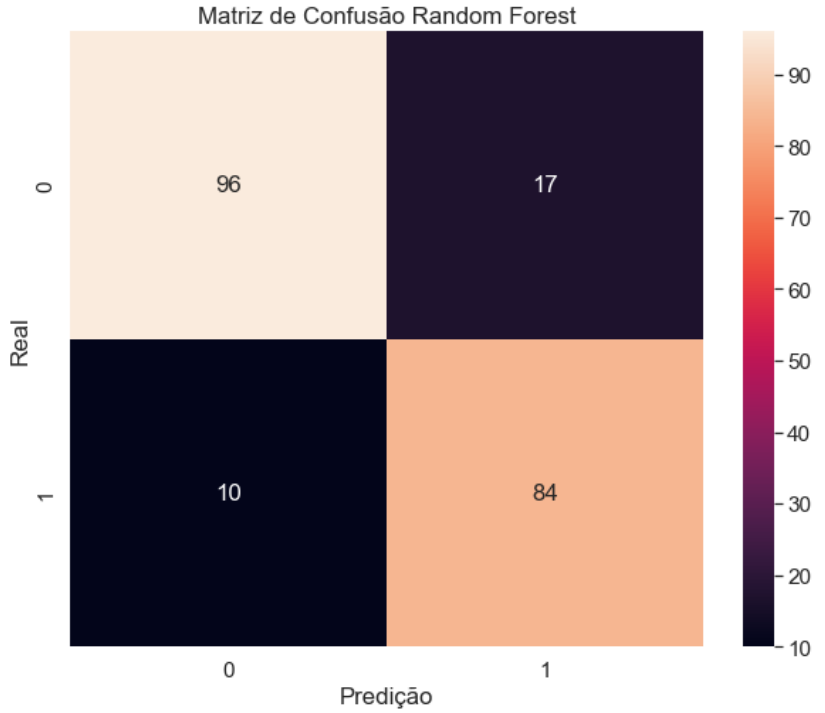

Figura 10: Matriz de confusão Random Forest

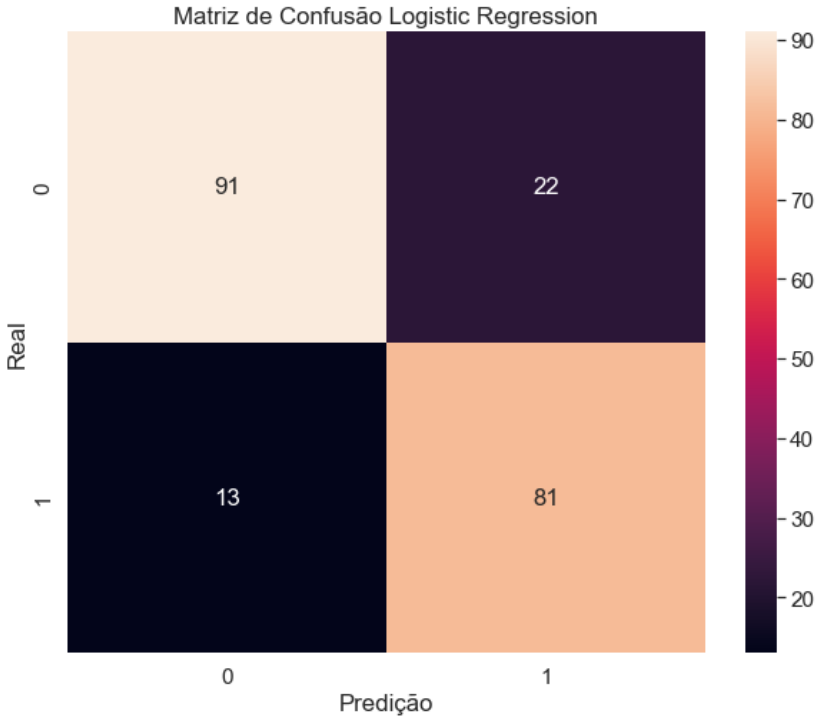

Figura 11: Matriz de confusão Logistic Regression

7. Conclusão

Para o projeto, o modelo escolhido foi o Random Forest. Pois apresenta uma maior acurácia e com uma matriz de confusão com os valores errados (Falso Positivo e Falso negativo) em menor quantidade.

Outro ponto importante a ressaltar é a análise do Cross Validation. O Random Forest alcançou 87.3% enquanto o outro modelo 87,13%. Lembrando que a acurácia da Logistic foi de 83 %. Sendo assim, há uma grande diferença entre as acurácias. Já o modelo Random Forest tem uma menor diferença, (87,3 - 87).