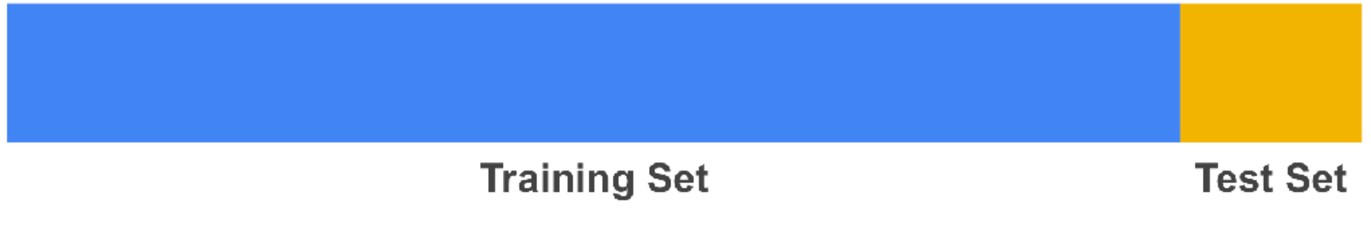


https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

Test Set

O volume de dados é suficientemente grande para trazer resultados estatisticamente significativos

Representa o dataset como um todo (desbalanceamento)

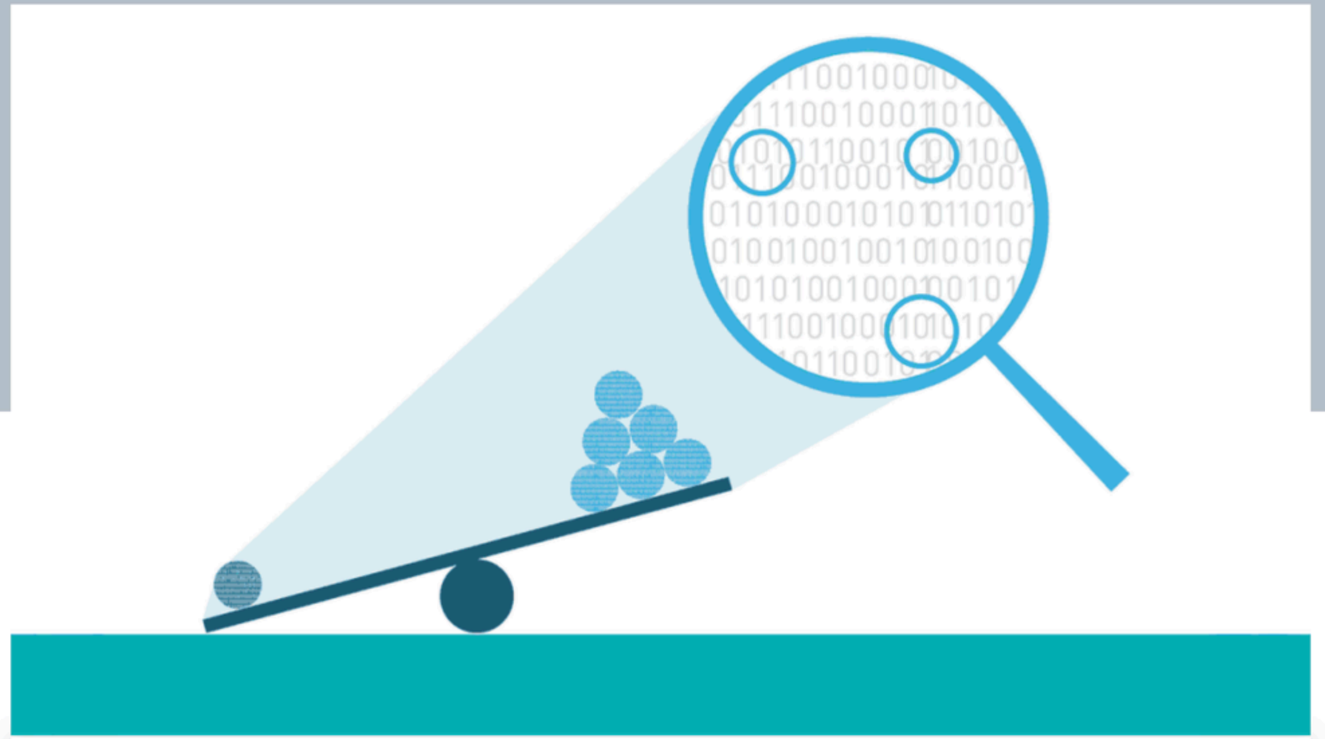
SIGMOIDAL

Blog Artigos eBook Gratuito Sobre Contato

BLOG

Como lidar com dados desbalanceados?

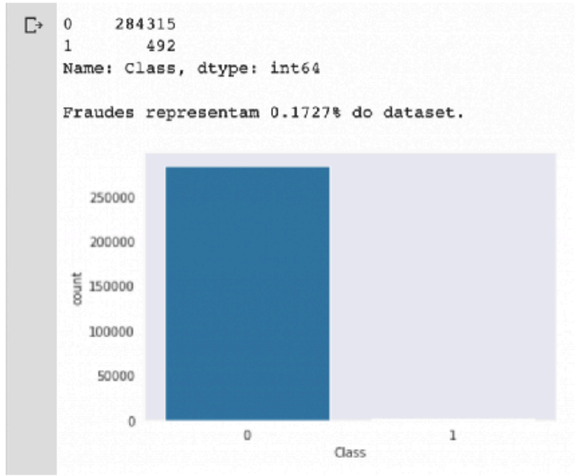
Escrito por Carlos Melo
em 24/12/2019



dados. Mal conseguimos ver a barra da nossa variável alvo para instâncias de fraude (df.Class == 1).

```
file_path = "https://www.dropbox.com/s/b44o3t3ehmnx2b7/creditcard.csv?dl=1"

# importar os dados para um dataframe
df = pd.read_csv(file_path)

# ver o balanceamento das classes
print(df.Class.value_counts())
print("\nFraudes representam {:.4f}% do dataset.\n".format((df[df.Class == 1].shape[0] / df.shape[0]) *

# plotar gráfico de barras para as Classes
sns.countplot('Class', data=df);
```

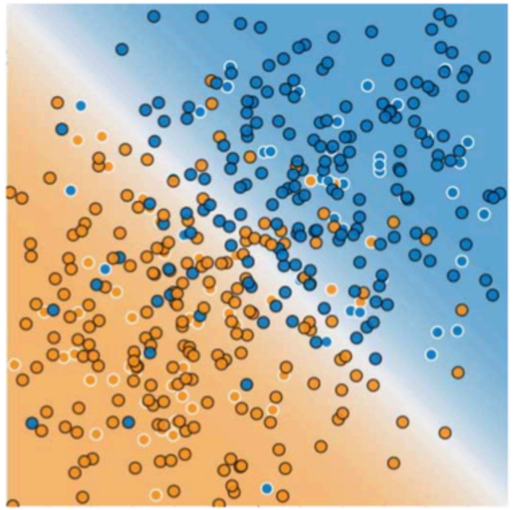
```
0    284315
1       492
Name: Class, dtype: int64

Fraudes representam 0.1727% do dataset.
```



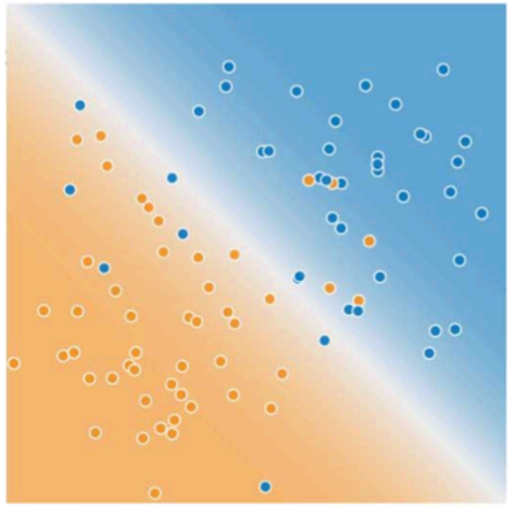
Para você ter um ideia real das consequências dessa situação, vou construir dois modelos de Regressão Logística. No nosso primeiro modelo, separei as variáveis X e y normalmente e dividi entre conjuntos de treino e teste, como é praxe em *machine learning*.

Sem maiores ajustes, treinei o modelo usando o método fit(X_train, y_train) e fiz a previsão de valores em cima do conjunto de teste (X_test). Na sequência, plotei a matriz de confusão e o relatório de classificação.

```
# separar variáveis entre X e y
X = df.drop('Class', axis=1)
y = df['Class']
```



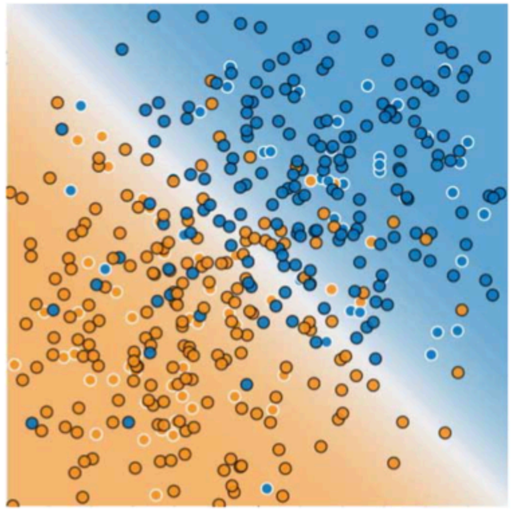
Training Data



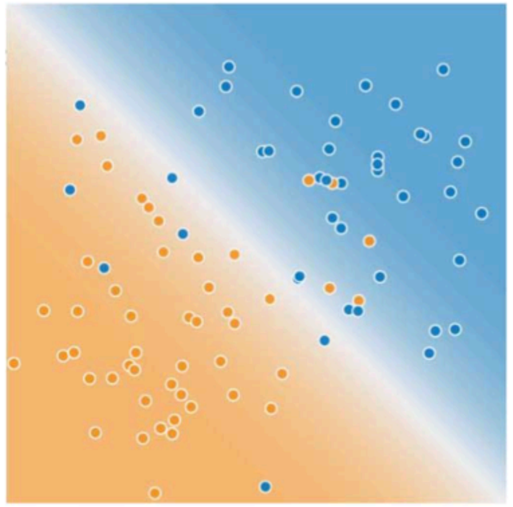
Test Data

https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

Um bom modelo tem desempenho bom nos datasets de treino e teste



Training Data



Test Data

https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

NUNCA treine seu modelo nos dados de treino.

Acurácia alta é um indicativo de overfitting. Verifique seus dados de teste.