www.datascience.pe

TEMA: Clasificación con datos desbalanceados

ESTRATEGIAS DE MUESTREO

Agenda

Estrategia

- Reto
- Métricas
- Muestreo (Sampling)
- Ajuste de Parámetros (Tuning)

Alerta

https://github.com/luish910/DSRPeru_imblearn

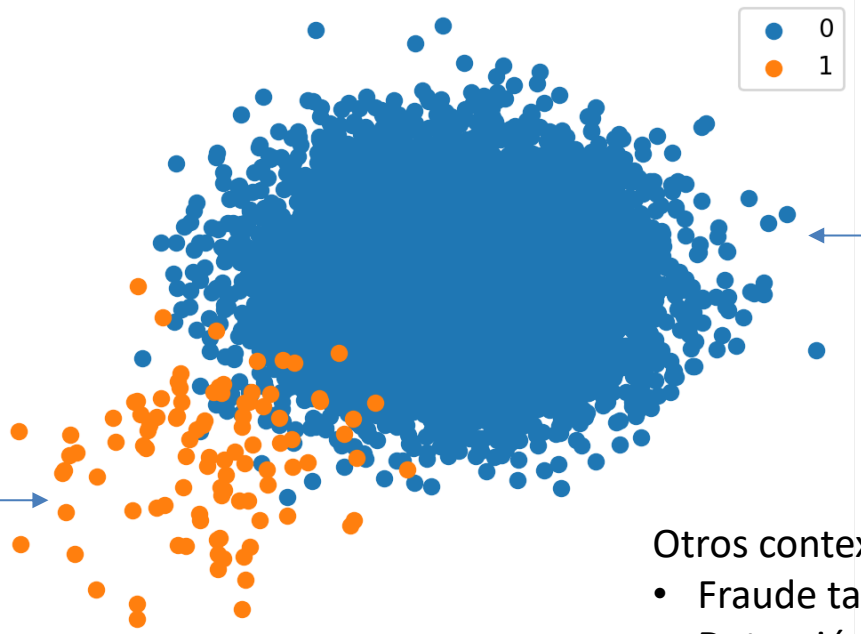
Data Science

…Somos un banco y enviamos un email a 12,500 clientes invitándolos a abrir un préstamo

¿Cuántos (%) responderán positivamente a la campaña?

Reto

N = 12,500 clientes



99%
No respondieron a la campaña

1%
Abrieron una préstamo

Otros contextos:

- Fraude tarjeta de crédito
- Detección COVID-19
- Churn (idealmente)

La trampa del Accuracy (Exactitud)

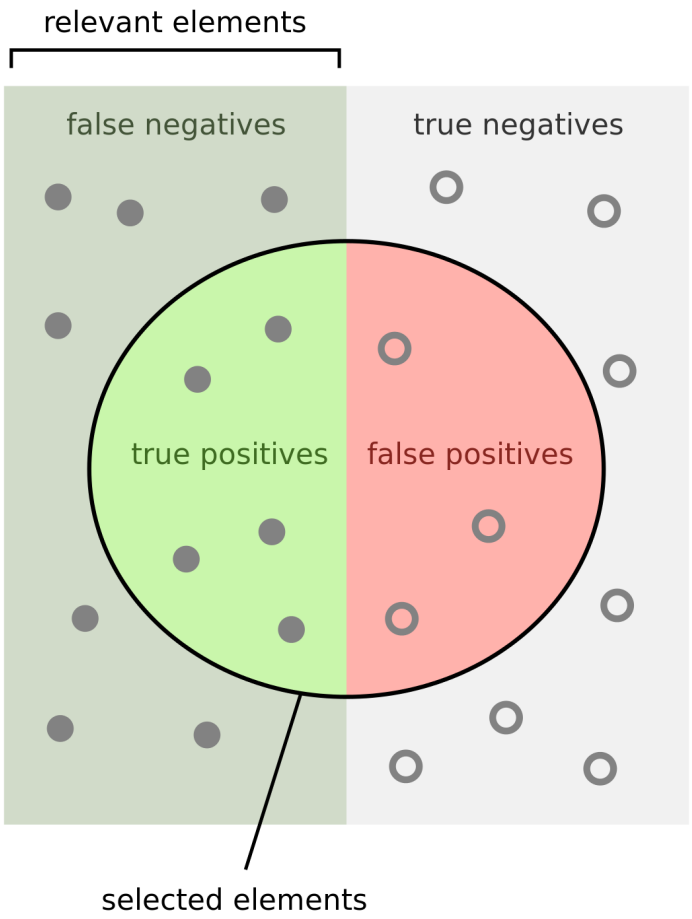
N = 12,500	Predicción 0	Predicción 1
Realidad 0	12,300	75
Realidad 1	115	10

Alerta

Accuracy = 98.4%

Un modelo puede tener un alto *accuracy* a pesar de ser deficiente prediciendo la clase menor o “rara”, o “relevante”.

Métricas



How many selected items are relevant?

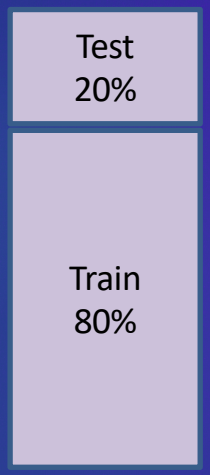
$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

- Precision: si el costo de FP es alto
- Recall: si el costo de FN es alto
- F1 Score: balance entre Precision y Recall

Dividir Train / Test sets



2,500 clientes

10,000 clientes

Para entrenar un modelo de Machine Learning, dividimos el conjunto de datos inicial en 2: entrenamiento (train) y pruebas (test).

El reto original (99% vs 1%) debe mantenerse en ambos sets

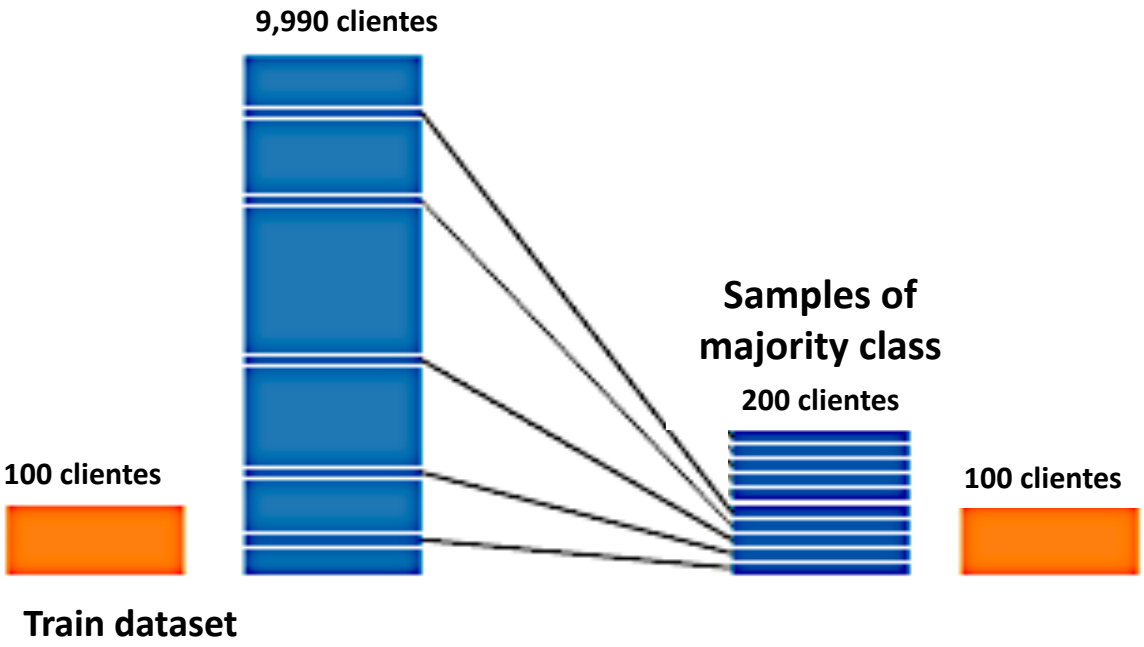
Baseline Performance

Usando Recall como métrica

Algoritmo	Recall	Desviación Estándard
Regresión Logística	49%	8.0%
Support Vector Machine (SVM)	43%	6.8%
Random Forest	63%	8.7%
Red Neuronal - ANN	45%	11.4%

Nuestro objetivo es aplicar técnicas de sobremuestreo y submuestreo para mejorar las métricas base

Random Undersampling



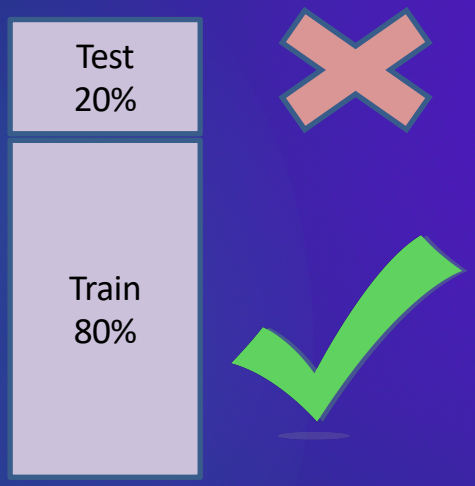
En nuestro ejemplo, reduciremos la clase mayor (0) para obtener un ratio de 2:1.

Random Undersampling

Performance

Algoritmo	Recall	Desviación Estándar
Regresión Logística	83%	9.8%
Support Vector Machine (SVM)	82%	6.8%
Random Forest	85%	12.2%
Red Neuronal - ANN	49%	40.0%

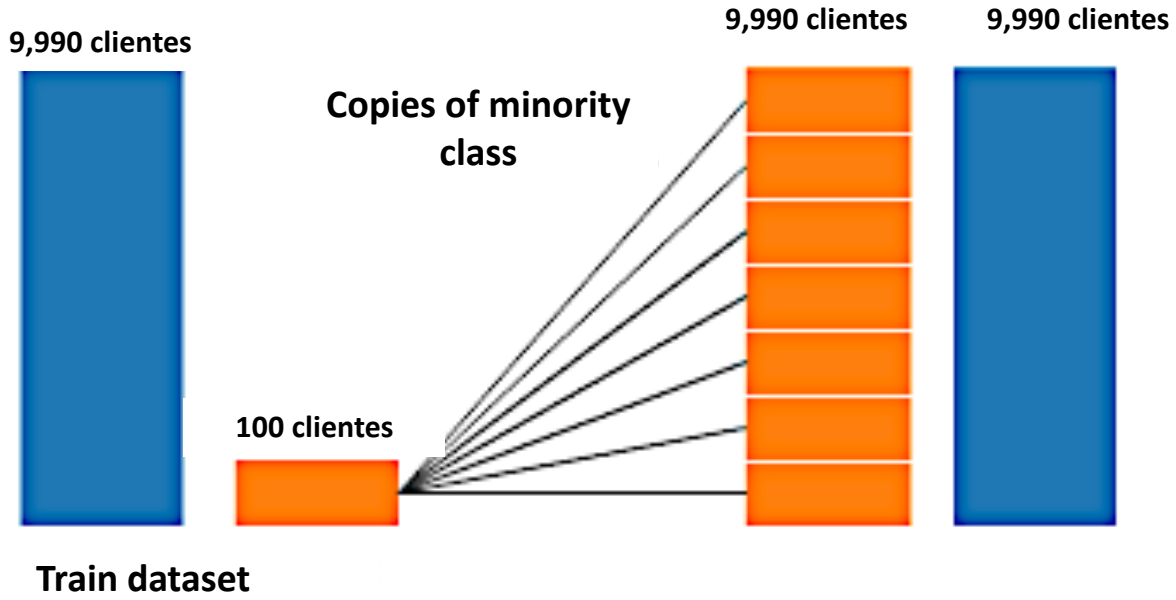
No muestrear la data de pruebas (test set)!



Alerta

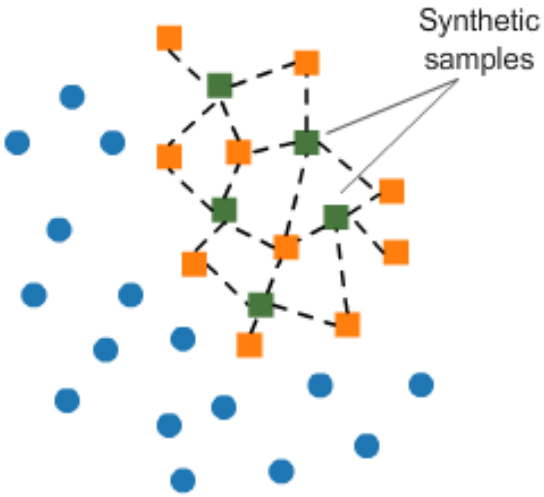
Solo aplicar la técnica de muestreo a la data de entrenamiento (train).

Random Oversampling



SMOTE Oversampling

Original dataset



Transformed dataset with synthetic samples

En nuestro ejemplo, crearemos nuevas observaciones para la clase menor (1) para obtener un ratio de 2:1.

SMOTE Oversampling

Performance

Algoritmo	Recall	Desviación Estándar
Regresión Logística	83%	8.7%
Support Vector Machine (SVM)	83%	8.7%
Random Forest	79%	11.1%
Red Neuronal - ANN	79%	3.7%

Data Science

…Si ya aplicamos Undersampling y Oversampling

¿Qué podemos aplicar ahora?

SMOTE + Random Undersampling

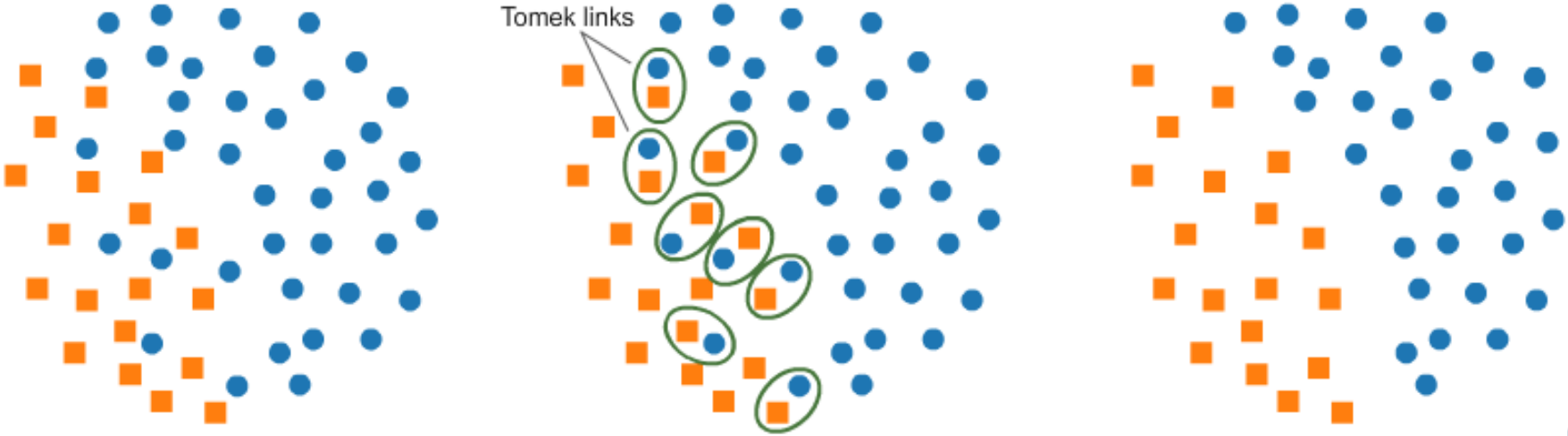
Performance

Algoritmo	Recall	Desviación Estándar
Regresión Logística	87%	8.7%
Support Vector Machine (SVM)	88%	7.5%
Random Forest	86%	11.6%
Red Neuronal - ANN	87%	8.7%

En general, se observa un incremento en performance en todos los algoritmos

Tomek Links

Orginal dataset



En nuestro ejemplo, aplicaremos SMOTE (oversampling) junto a Tomek Links (undesampling solo a la clase mayor – 0) obteniendo un ratio de 1:1.

SMOTE + Tomek Links

Performance

Algoritmo	Recall	Desviación Estándar
Regresión Logística	87%	7.5%
Support Vector Machine (SVM)	89%	8.0%
Random Forest	82%	7.5%
Red Neuronal - ANN	78%	5.8%

Para este dataset, la combinación solo mejoró el performance de los 2 primeros algoritmos

Data Science

Performance: Comparación

	Baseline	Random Undersampling	SMOTE	SMOTE + Random Und.	SMOTE + Tomek L.
Algorithym	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**
Logistic Regression	49%	83%	83%	87%	87%
Support Vector Machine	43%	82%	83%	88%	89%
Random Forest	63%	85%	79%	86%	82%
Artificial Neural Network	45%	49%	79%	87%	78%

Grid Search

Todos los algoritmos incluyen un conjunto de hiperparámetros que nos permiten controlar su comportamiento. Estos deben ser fijados antes del entrenamiento.

La creación de “pipelines” nos permiten aplicar las técnicas de muestreo aprendidas dentro de un “Grid Search” o buscador de parámetros óptimos.

Ajuste de Parámetros (Tuning)

Seleccionamos las top 5 combinaciones de algoritmo + técnica de muestreo. Con ellas podemos buscar el ajuste de parámetros

	SMOTE + Random Und.	SMOTE + Tomek L.
Algorithym	**Recall**	**Recall**
Logistic Regression	87%	87%
Support Vector Machine	88%	89%
Random Forest	86%	82%
Artificial Neural Network	87%	78%

SVM (SMOTE + Tomek Links):
Se obtuvo un recall de **88%** con los parámetros Kernel = linear , c = 50

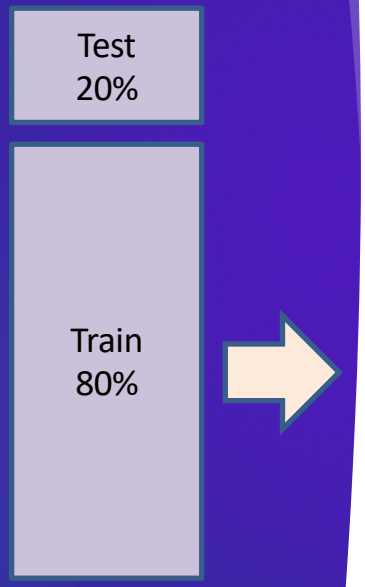
Random Forest (SMOTE + Random Undersampling):
Se obtuvo un recall de **92%** con los parámetros n_estimaros = 50 , max_depth = 6

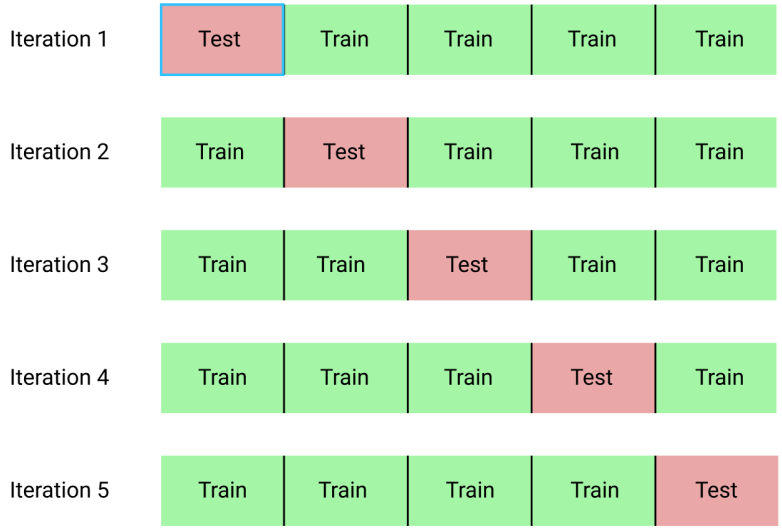
No tocar el Test set para el ajuste de parámetros

Alerta

ANEXO:
Cross-Validation
(Validación cruzada)

El proceso de validación cruzada es repetido durante k=5 iteraciones, con cada uno de los posibles subconjuntos de datos de prueba





Los “pipelines” también nos permiten aplicar validación cruzada de tal manera que el set de prueba (test) se mantenga intacto.

Estrategia

Seleccionar métrica

- Precision, Recall
- F1: Balance entre Precision y Recall
- Accuracy (!): ambas clases son importantes

Entrenar modelos base

- Parámetros comunes

Muestreo

- Undersampling: Random, Tomek Links
- Oversampling: Random, SMOTE
- Combinado

Ajuste de Parámetros

- Grid Search: vía pipelines

Entrenar todas las combinaciones posibles

Top 5 combinaciones

Contacto

Luis H. Murrugarra

luish910@gmail.com

linkedin.com/in/lmurrugarra

Fuentes

- https://www.researchgate.net/publication/328315720_Cross-Validation_for_Imbalanced_Datasets_Avoiding_Overoptimistic_and_Overfitting_Approaches
- https://kiwidamien.github.io/how-to-do-cross-validation-when-upsampling-data.html
- https://gist.github.com/kiwidamien/bcbe8e527a5f0cc9f28c4fe692f70cbc
- https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/
- https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
- https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets
- https://www.udemy.com/share/101YDS/
- https://towardsdatascience.com/what-to-do-when-your-classification-dataset-is-imbalanced-6af031b12a36
- https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Data Science

www.datascience.pe