Modelar sobre plataforma de ML

Programa Ejecutivo IA & Deep Learning

Profesor:
Andrés González García - CTO y Socio Cofundador de CleverData.io
agonzalez@cleverdata.io

www.eoi.es

Índice

Recursos recomendados

Artículos

- Una introducción visual al ML. Explicación muy visual y didáctica de cómo funcionan los árboles de decisión. La mejor que he visto nunca.

- Model Tuning and the Bias-Variance Tradeoff. Segunda parte del artículo anterior que explica dos de los problemas clásicos de lo modelos de Machine Learning: el *bias* (modelos demasiado simples) y la *varianza* (también llamado overfitting, modelos demasiado complejos).

- Preguntas frecuentes sobre Machine Learning y BigML.

- Cinco consejos para empezar con Machine Learning en la empresa. Basado en una charla del director de ML de Uber.

- Tutoriales de BigML no escritos por BigML (bueno, algunos sí).

- Artículos introductorios al Machine Learning.

Recursos recomendados

Vídeos

- Vídeo tutoriales de BigML. Es un buen compendio de vídeos cortos. Si tienes dudas sobre algún concepto o cómo usar algo de la plataforma, no dudes en pasarte por aquí: modelos, ensembles, evaluaciones, clusters, sources, datasets… hay de todo. En inglés.

Una consideración previa

Antes de empezar con los ejercicios quería remarcar que es muy importante entender los datos antes de crear los modelos. Los datos que vamos a usar en estos ejercicios son prácticamente autoexplicativos. Es decir, no necesitan una documentación que los explique, sino que los nombres de las variables ya dicen lo que son.

Para hacer estos ejercicios no es estrictamente necesario entenderlos, porque además de ser autoexplicativos, son datos muy preparados y que dan buenos resultados con los modelos con los que vais a trabajar. Los problemas llegan cuando los modelos no funcionan y hay que entender por qué. Y si no sabemos qué nos están contando los datos, lo tendremos muy muy complicado. Pero eso no va a suceder en este curso.

Mi recomendación es que antes de poneros con los modelos, descarguéis los datos en vuestros ordenadores y los visualicéis con Excel o con cualquier otra herramienta que hayáis estudiado ya, como Tableau. No lo he puesto explícitamente en cada ejercicio, pero **es muy deseable que lo hagáis**. Veréis también que en la plataforma BigML se generan automáticamente unos histogramas que ayudan bastante.

En el mundo real… lo siento, pero los datos están bastante “sucios” y en numerosas ocasiones no encontraréis a quien realmente sabe lo que significan. Hay que captarlos, entenderlos, limpiarlos, transformarlos y seleccionar las variables que más van a servir para el objetivo de cada proyecto. Estas fases son más importantes que las propias de Machine Learning.

¡Suerte!

Índice

- Recursos recomendados.
- **Ejercicio 4. Detector de anomalías.**

Ejercicio 4. Detector de anomalías

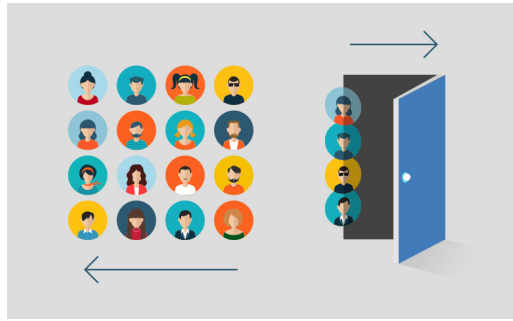
DIY!

- Objetivo: usar Machine Learning para mejorar los modelos de Machine Learning. En este ejercicio usarás el detector de anomalías para limpiar un dataset y crearás tu primer modelo de tipo ensemble.

- Supuesto: eres el responsable de marketing de una empresa de telefonía. Quieres crear un modelo con Machine Learning que te diga qué clientes están en peligro de darse de baja de tu compañía para poder hacerles una oferta a la que no pueden negarse.

- Haz un RANDOM SPLIT 80/20 y crea un **modelo de tipo ensemble 1-CLICK ENSEMBLE**.

- Crea la evaluación con el 20% del dataset de test.



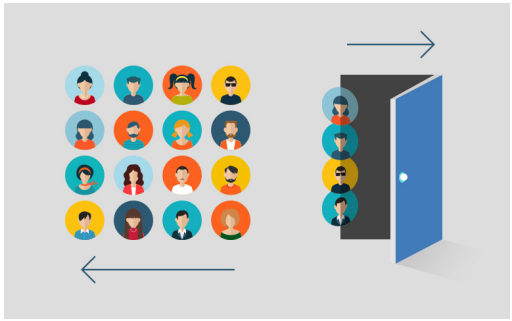
https://cleverdata.io/csv/churn-telecom.csv

Ejercicio 4. Detector de anomalías

DIY!

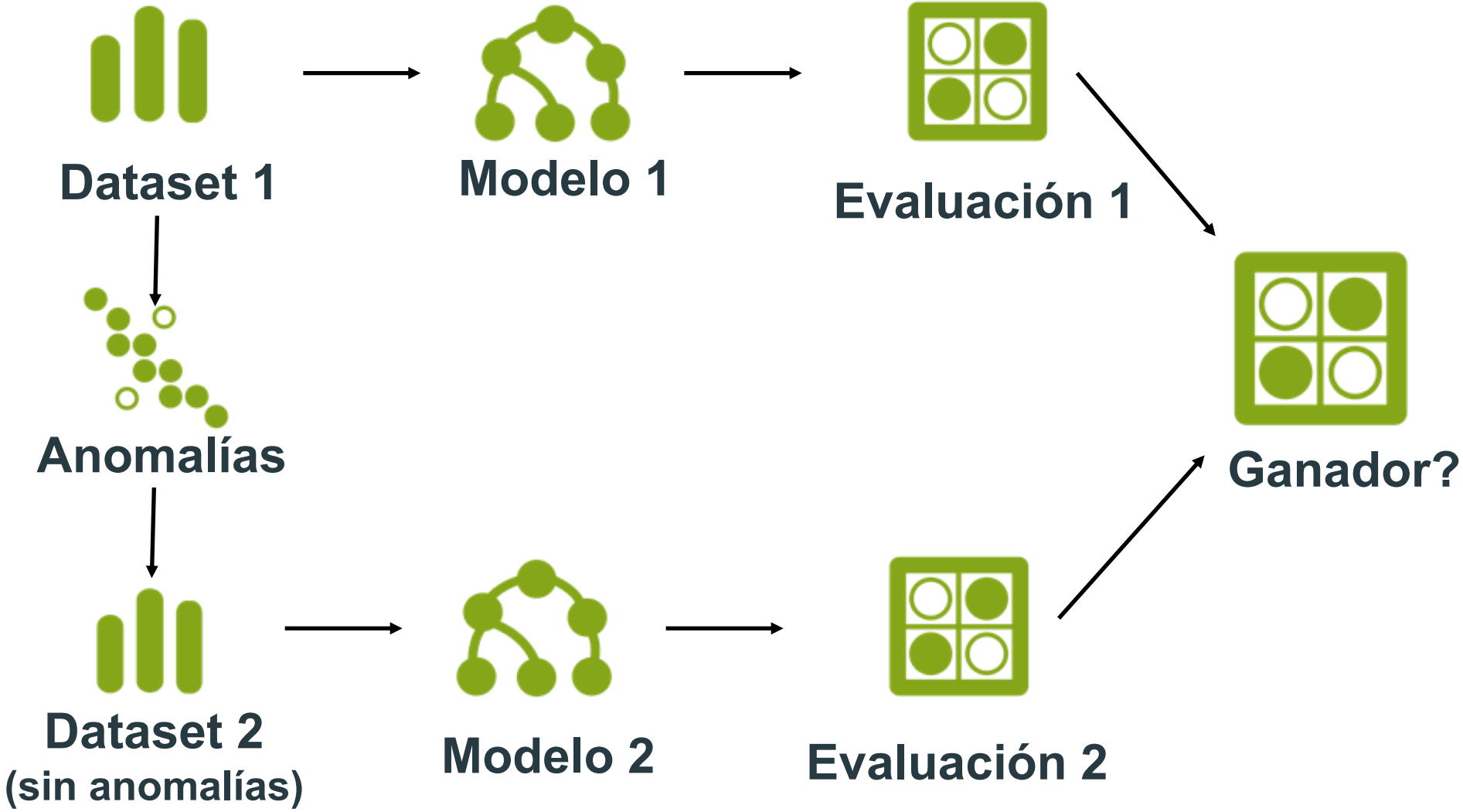
- Vuelve al dataset del 100% de los datos y detecta las 10 primeras anomalías.
- Crea un nuevo dataset a partir de este eliminando las 10 primeras anomalías.
- Crea un nuevo Split 80/20 y un nuevo **ensemble** con este nuevo dataset.
- Haz la **evaluación** con el 20%.
- **PREGUNTA 4.1**: Compara los resultados de calidad del modelo. ¿Cuál es mejor?

	False Modelo 1	True Modelo 1	False Modelo 2	True Modelo 2
Accuracy				
Precision				
Recall				
PHI Coefficient				



- Material de apoyo.
  - Vídeo: Detección de anomalías
  - Artículo: How to use Machine Learning to improve your Machine Learning models

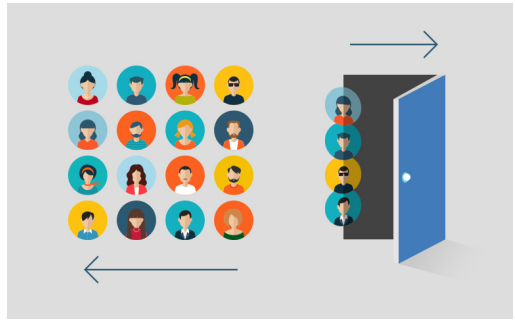
Ejercicio 4. Detector de anomalías. Workflow



Ejercicio 4. Detector de anomalías

DIY!

- **Enlaces a entregar**:
  - Dataset:
  - Dataset 1 80%:
  - Dataset 1 20%:
  - Dataset 2 80%:
  - Dataset 2 20%:
  - Ensemble 1 80%:
  - Ensemble 2 80%:
  - Evaluación 1:
  - Evaluación 2:



agonzalez@cleverdata.io