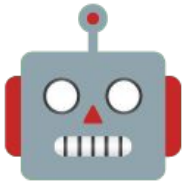
AutoML



AutoML

¿ Qué es el AutoML ? ¿ Para qué sirve ?

Hasta el momento hemos discutido muchos temas sobre de Aprendizaje de Máquina, esta vez vamos a aprender sobre el Aprendizaje Automático (Auto-ML) y cómo apoya a los ingenieros de ML para acelerar el desarrollo de ML y ahorrar los esfuerzos y el ahorro de costes a los clientes.

La idea del AutoML es poder tener de manera automatizada todo el proceso que venimos viendo hasta ahora en el curso normalmente se considera que un sistema de AutoML debe contar de las siguientes funcionalidades:

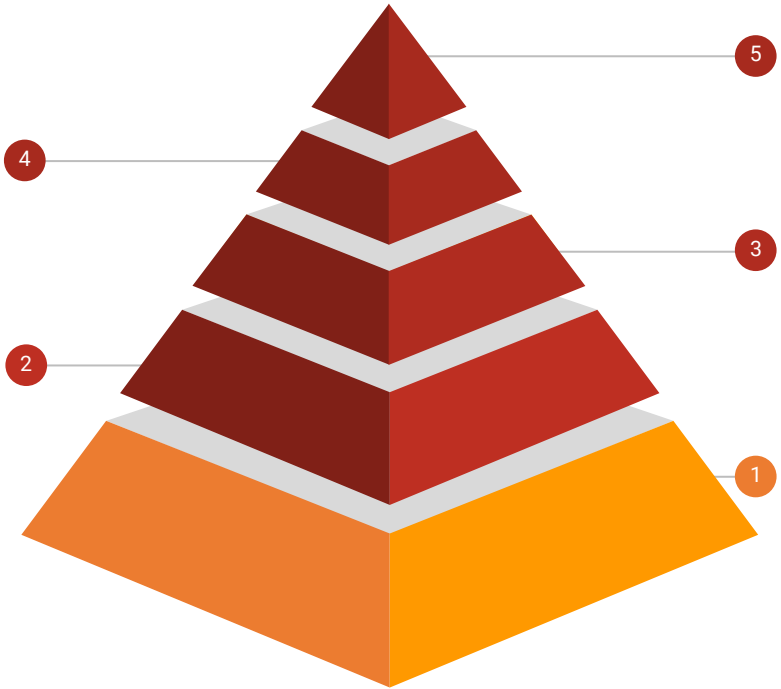
AutoML – Pipeline (proceso)

Evaluación

Evaluar diferentes tipos de modelos en busca del que más se ajuste a nuestro problema

Limpieza

Limpieza de datos e ingeniería de características.



Despliegue

Despliegue del modelo para su rápida utilización-

Construcción de modelos

Construcción de modelos para seleccionar el algoritmo de ML correcto.

Recolección (No incluido)

Recolección de la data necesaria desde distintas fuentes ya sean archivos sueltos o extracción desde base de datos.

Impacto de AutoML en los proyectos de ciencia de datos

Comprensión del negocio

El primer paso en cualquier proyecto de ciencia de datos es entender bien el problema. En general, AutoML es el que menos impacto tiene en esta área. Lo mejor que puede hacer AutoML es automatizar algunos proyectos de ciencia de datos populares y estándar. Un ejemplo es la predicción del fraude bancario. Los patrones de fraude pueden no cambiar mucho de un cliente a otro. Será fácil crear un modelo de solución y llevarlo al mercado.

¿Automatizará AutoML todos los problemas empresariales?

La respuesta es definitivamente no. Hay muchos escenarios que pueden ser automatizados o incluso replicados pero hay muchos otros que no y que requieren del conocimiento del negocio.

Impacto de AutoML en los proyectos de ciencia de datos

Recolección de datos

Aunque existen productos de AutoML que permiten incorporar nuevos datos directamente al proceso, por ejemplo con GCP es fácil importar datos a las tablas de AutoML como archivos planos o utilizando BigQueries.

La parte de recolección de los datos está muy ligada al trabajo del científico de datos ya que este debe conocer el problema que se quiere atacar y realizar la extracción de los datos correspondientes que sean de interés para el mismo.

Impacto de AutoML en los proyectos de ciencia de datos

Limpieza de datos

Tras importar el conjunto de datos necesario, el siguiente paso es limpiar los datos. Este paso suele ser tedioso y requiere mucha atención por parte de los científicos de datos. Los conjuntos de datos no suelen estar lo suficientemente limpios para el consumo de los modelos ML. Las soluciones AutoML serán muy útiles en este caso. Podremos limpiar los datos mucho más rápido.

AutoML será capaz de acelerar la limpieza de datos. Será posible llevar el esfuerzo requerido de semanas a días. Pero, el conocimiento del dominio de un científico de datos será la clave para alcanzar la mejor solución.

Impacto de AutoML en los proyectos de ciencia de datos

Ingeniería de características

Se trata de un proceso iterativo y una de las etapas que más tiempo consumen en un proyecto de ciencia de datos. Con AutoML será fácil implementar algunas tareas de ingeniería de características. Tareas como la normalización, la codificación, etc pueden realizarse con un clic.

¿Se puede automatizar completamente la ingeniería de características?

La respuesta corta es no. Muchas de las tareas realizadas en la ingeniería de características serán accesibles. Sólo el equipo de ciencia de datos puede incorporar las ideas de negocio en las características. Utilizar las características y transformaciones conocidas sólo ayudará hasta cierto punto. Para lograr un alto rendimiento, hay que realizar una exploración más profunda.

Impacto de AutoML en los proyectos de ciencia de datos

Construcción de modelos

El AutoML nos va a permitir poder probar muchos modelos de manera rápida haciendo que la selección, el ajuste y el seguimiento del modelo sean fáciles de implementar. De este modo, se genera tiempo para que el equipo de ciencia de datos trabaje en más problemas.

Esto también nos permitirá focalizar nuestros esfuerzos a otras áreas que necesitan mucho conocimiento del negocio permitiendo que los equipos de ciencia de datos se encarguen de más tareas basadas en la información y como extraer más valor de las mismas.

Impacto de AutoML en los proyectos de ciencia de datos

Evaluación y seguimiento

Este es un aporte muy importante del AutoML ya que nos permite poder evaluar diferentes modelos considerando el criterio o métrica que nosotros deseemos pudiendo tener un registro completo de los resultados de cada uno de los mismos, de manera de tener un seguimiento que nos permite elegir luego los que mejor resultados nos den.

Impacto de AutoML en los proyectos de ciencia de datos

Deployment (despliegue)

En muchos proyectos de ciencia de datos, el despliegue nunca ha sido fácil. Hay problemas a la hora de trasladar los modelos de un entorno a otro. Cualquier pequeña diferencia entre los entornos, como las versiones de software, podría causar problemas. Además, el entorno de producción suele ser restringido. Esto dificulta la realización de cambios o el seguimiento del rendimiento de los modelos ML.

AutoML nos agiliza mucho este proceso permitiéndonos desplegar modelos en cuestión de minutos.

AutoML Ventajas y Desventajas

Ventajas

- Podemos utilizarlo tanto para problemas supervisados como no supervisados.
- Podemos utilizarlo tanto para regresión como para clasificación.
- Ahorro de tiempo.
- Automatización de tareas repetitivas y manuales que son susceptibles de errores humanos.
- Facilita la selección de modelos y la supervisión del rendimiento
- El ajuste de los hiperparámetros puede automatizarse completamente con AutoML

AutoML Ventajas y Desventajas

Desventajas

- Pérdida de visión global del problema.
- Si no se conocen los algoritmos que se están utilizando puede correrse el riesgo de usarse como una caja negra.

AutoML

¿ Quienes nos ofrecen este servicio ?

Y muchos mas...

AutoML

Nosotros elegimos mostrar en la cátedra:

Con integración de:

Instalación:

```
pip install pycaret
pip install mlflow
```

Aunque los invitamos a probar otras!

AutoML

Nuestra sugerencia se basa en lo siguiente:

- Pycaret se ha creado pensando en la velocidad y la automatización - Pycaret se autodefine como una biblioteca de aprendizaje automático de código abierto en Python que permite pasar de la preparación de los datos al despliegue del modelo en cuestión de minutos (y cuando dicen minutos, quieren decir minutos). Permite conseguir un primer modelo entrenado en unos pocos minutos y ejecuta todos los modelos de scikit-learn.

- Permite el uso GPU!

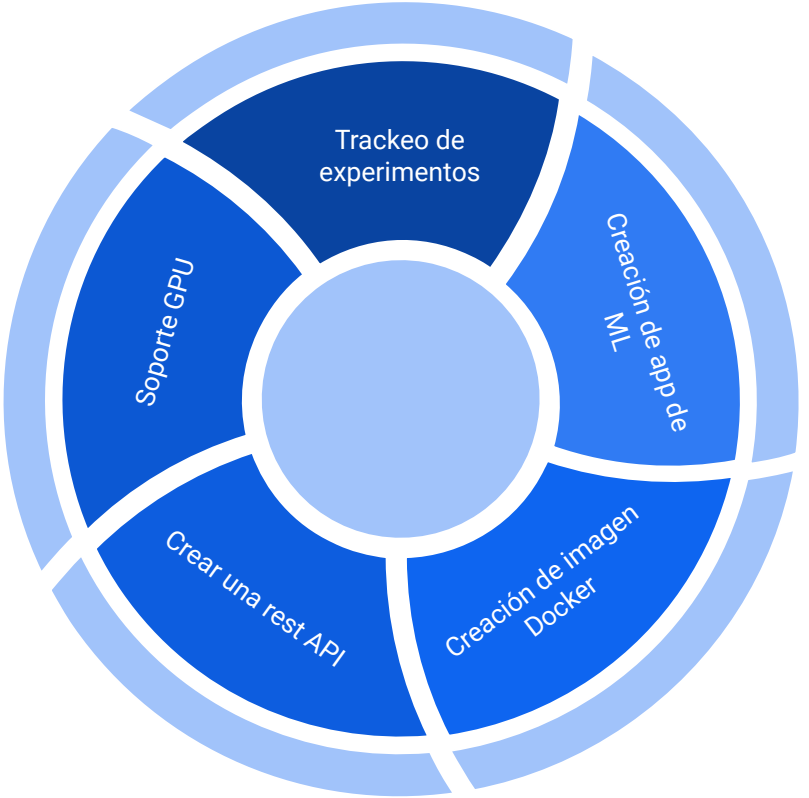
AutoML

Nuestra sugerencia se basa en lo siguiente:

- Lo que le falta a Pycaret en el número de funciones generales, lo compensa con la simplicidad. La función setup() realiza toda la preparación de los datos, desde la codificación de las categorías hasta la imputación de los valores perdidos (missing) , la eliminación de los valores atípicos, etc. Todo esto se hace en una sola función en lugar de repartir el trabajo en una serie de llamadas a funciones diferentes y difíciles de recordar. La función compare_models() ejecuta un benchmark contra todos los algoritmos aplicables y devuelve los datos de rendimiento.

- Evaluación y despliegue sencillos - Funciones como evaluate_model() e interpret_model() devuelven interfaces fáciles de usar para desarrollar una comprensión más profunda de su modelo. La función deploy_model() permite al usuario un proceso para desplegar modelos en AWS, Google Cloud o Azure. Los usuarios también pueden crear un objeto compatible con scikit-learn para el despliegue en otros entornos.

AutoML


Trackeo de experimentos
Creación de app de ML
Creación de imagen Docker
Crear una rest API
Soporte GPU

Bibliografía

- Pycaret
- Azure AutoML
- Google AutoML
- Amazon AutoML