MeLi Challenge

Matías Sorozábal

28-06-2022

Solución propuesta:

Inicialmente al ver el juego de datos se me ocurrió hacer 2 cosas:

1. Vi que habían URLs de publicaciones, por lo cual me puse a ver si podíamos hacer get de esas imgs y construir una CNN y quizas combinarla con Unet o YOLO object detection para clasificar las imágenes .

   claramente no tengo un dataset para entrenar con imágenes de artículos usados y nuevos, pero podría ser interesante explorar este approach.

2. Un clasificador que me permita trabajar con datos categóricos. *(ver meli-challenge.ipynb)*

Se realizó un trabajo de limpieza (parseo) de datos para llevarlo a un esquema tabular.

Luego un EDA para entender la distribución de algunos datos, correlaciones entre los campos, completitud, schemas, análisis de outliers.

Se construyó un modelo baseline que utiliza el algoritmo de CatboostClassifier con earlystopping para evitar overfitting.

La métrica adicional elegida es el f1-score (media armónica entre el precision y el recall), se agrega el plot del threshold y me parece que puede ser interesante porque nos va a permitir calibrar en función al negocio que es más importante, si agarrarlos a todos (recall) y no errarle (precision).

Además se visualiza la ROC AUC que nos muestra que de manera muy efectiva el modelo logra distinguir las clases. (usado y nuevo)

Se exporta el artefacto.
Mejoras:

- Realizar un mejor trabajo de ingeniería de features.. además agregar las columnas que se eliminaron por tener dicts adentros que pueden tener información relevante pero para este baseline quedaron afuera.

- Si bien se agregó el código del Grid Search no fue realizado por cuestiones de cómputo y que el modelo ya superó el mínimo en su versión baseline.

- Disponibilizar el modelo en una API para recibir invocaciones a través de un endpoint.

- Se podría evaluar la posibilidad de generar el modelo de Computer vision mencionado y tener una especie de ensamble de modelos para sumar insight a la toma de decisión.
- Generar Pipeline de entrenamiento y deploy.