30 days of ML

11 Agosto 2021

Validación de modelos (más comunes)

- Holdout esquema:

- Clase 4, Intro ML: *train_test_split*
- *StratifiedShuffleSplit(n_split = 1, test_size = 0.2, random_state = 42)*

- K-fold esquema:

- *Kfold*
- *TimeSeriesSplit*
- *StratifiedKfold*

- LOO (leave-one-out):

comparación scikitlearn

Cómo hacer el split

1. **Random** / Por filas: Suponemos que todas las filas son independientes

2. **Temporal**: Rossmann Store Sales, Grupo Bimbo Demanda

3. Por **ID**: Diferentes usuarios en train y test -> No podemos aplicar features basadas en ID. (Ej. CAT)
   obs: La ID puede estar oculta.(ej. Intel & MobileODT Cervical Cancer Screening)

4. **Mix:** Ej. Predecir ventas de distintas tiendas. (ID y Temporal Split)

Obs: Validación mal hecha -> Overfitting
Tip: Tratar de repetir train/test split realizado por los organizadores.

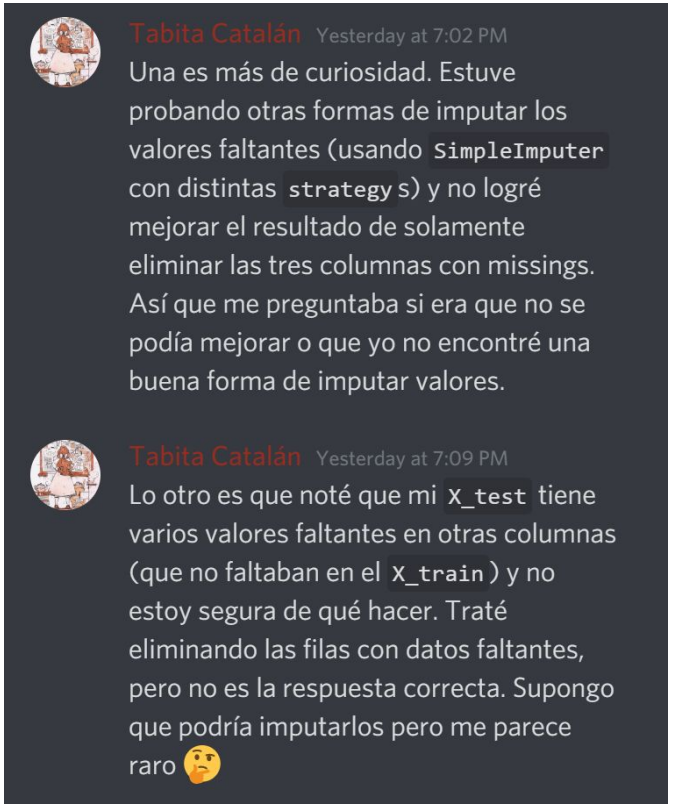
Leaderboard probing

1. Entregar predicciones constantes en problemas con datos tabulares.

   ¿Qué ganamos haciendo esto en problemas de Clasificación y Regresión?

2. Ejemplos: The "Perfect Score" Script, 2nd Place Competition

Discusión sobre imputación de datos: ML Intermediate



- Por la mediana de un grupo

```
df['LotFrontage'] = df.groupby('Neighborhood')['LotFrontage'].transform(lambda x: x.fillna(x.median()))
```

- Utilizando otros imputers de scikitlearn:
  Ej: KNNImputer

- Si el problema es una serie de tiempo:
  - Interpolación: Lineal, Medías moviles, etc.

- ¿Qué otras formas conocen?
- ¿Qué hacen cuando el test set tiene ese problema?

Tópicos siguiente sesión

- Ensamblaje de modelos:

  - Stacking
  - StackNet (Stacking usando NN)
  - Blending
  - Bagging
  - Boosting

- Data Leakages

- ¿Otros temas?

¿Preguntas / Comentarios ?

Gracias por su tiempo.