
30 days of ML



TensorFlow

User Group Santiago

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



Tabita Catalán Yesterday at 7:02 PM

Una es más de curiosidad. Estuve probando otras formas de imputar los valores faltantes (usando `SimpleImputer` con distintas `strategy`s) y no logré mejorar el resultado de solamente eliminar las tres columnas con missings. Así que me preguntaba si era que no se podía mejorar o que yo no encontré una buena forma de imputar valores.



Tabita Catalán Yesterday at 7:09 PM

Lo otro es que noté que mi `x_test` tiene varios valores faltantes en otras columnas (que no faltaban en el `x_train`) y no estoy segura de qué hacer. Traté eliminando las filas con datos faltantes, pero no es la respuesta correcta. Supongo que podría imputarlos pero me parece raro 🤔

- Por la mediana de un grupo

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

- Utilizando otros imputers de [scikitlearn](https://scikit-learn.org/):
Ej: `KNNImputer`
- Si el problema es una serie de tiempo:
 - Interpolación: Lineal, Medías móviles, 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.**