Laboratorio de Implementación I

Edición 2023 Rosario

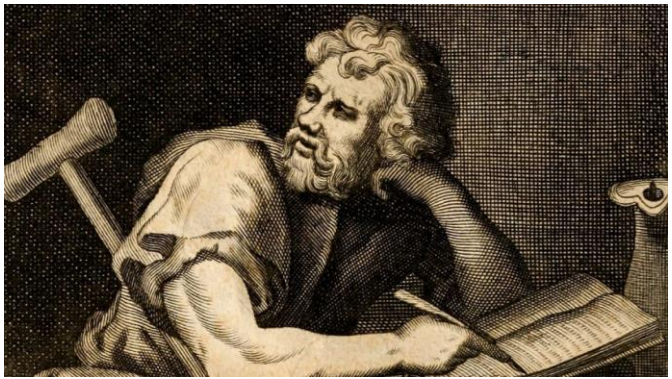
Maestría en Ciencia de Datos / Universidad Austral

Ing. Agr. Néstor Di Leo

Indice

- Introducción
- Metodología experimental
- Pruebas complementarias
- Resultados
- Conclusiones

“Los hombres no se perturban por las cosas, sino por la opinión que tienen de éstas” (Epícteto).



Introducción

- LABO 1: completo compendio de todo lo que tenga que ver con Modelos Predictivos…
  … y otras muchas cosas valiosas más…

- Tanto para EXPERIMENTOS COLABORATIVOS como para SEMILLERÍOS E HIBRIDACIÓN, el objetivo fue:

  Encontrar cuál es la mejor configuración de las etapas del workflow que genere la mayor ganancia en el Private Leaderboard de Kaggle.

Introducción

- Encontrar los mejores parámetros en LightGBM es esencial para obtener un modelo de aprendizaje automático con un rendimiento óptimo, evitar el sobreajuste y asegurar que el modelo generalice bien a nuevos datos.

- La selección adecuada de parámetros es una parte esencial del proceso de construcción y ajuste de modelos y puede tener un impacto significativo en el éxito del proyecto de aprendizaje automático.

Metodología experimental

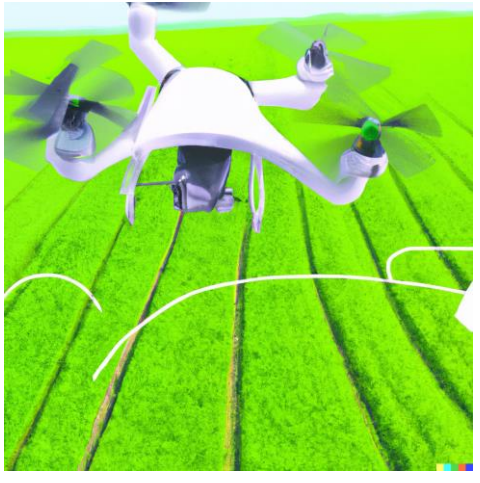
Experimento factorial :

2^4 →

- **RF: 20_5_500_100 / 25_6_200_200**
- **Canaritos: 0.5 y 0.5 / 0.66 y 1.0**
- **K-Folds en el Cross Validation: 5 / 10**
- **Undersampling s/ clase m+: 0.3 / 0.5**

Factores fijos:

- ML y Rank cero fijo
- Meses sin pandemia y Julio de 20221 en el training
- Semillerío final con **50** semillas
- Regularización fija en: 225; 325 y 4



Primeros resultados

Experimento	Cross Validation	Undersampling	Canaritos	Arbol	Regularización	Corrida VM	Ganancia Public
Testigo	5	0.4	0 y 0	20_4_1000_40			51.615
1	5	0.3	0.5 y 0.5	20_5_500_100	con regul	OK	46.739
2	10	0.3	0.5 y 0.5	20_5_500_100	con regul	OK	47.789
3	5	0.5	0.5 y 0.5	20_5_500_100	con regul	OK	46.351
4	10	0.5	0.5 y 0.5	20_5_500_100	con regul	OK	45.999
5	5	0.3	0.66 y 1.0	20_5_500_100	con regul	OK	47.053
6	10	0.3	0.66 y 1.0	20_5_500_100	con regul	impar	46.471
7	5	0.5	0.66 y 1.0	20_5_500_100	con regul	par	46.829
8	10	0.5	0.66 y 1.0	20_5_500_100	con regul	polska	46.719
9	5	0.3	0.5 y 0.5	25_6_200_200	con regul	impar	48.755
10	10	0.3	0.5 y 0.5	25_6_200_200	con regul	polska	47.025
11	5	0.5	0.5 y 0.5	25_6_200_200	con regul	par	48.289
12	10	0.5	0.5 y 0.5	25_6_200_200	con regul	polska	47.983
13	5	0.3	0.66 y 1.0	25_6_200_200	con regul	par	47.971
14	10	0.3	0.66 y 1.0	25_6_200_200	con regul	polska	48.189
15	5	0.5	0.66 y 1.0	25_6_200_200	con regul	impar	47.819
16	10	0.5	0.66 y 1.0	25_6_200_200	con regul	par	47.713

Primeros resultados

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       47.68638    0.70992  67.171 9.88e-16 ***
CrossV            -0.04795    0.05367  -0.894 0.390689
Undersampling     -1.43113    1.34163  -1.067 0.308954
Canaritos0.66 y 1.0 -0.02077    0.26833  -0.077 0.939691
Arbol25_6_200_200    1.22425    0.26833   4.563 0.000813 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5367 on 11 degrees of freedom
Multiple R-squared:  0.6742,    Adjusted R-squared:  0.5557
F-statistic:  5.69 on 4 and 11 DF,  p-value: 0.00987

> summary(anova1)
             Df Sum Sq Mean Sq F value   Pr(>F)
CrossV        1  0.230   0.230   0.798 0.390689
Undersampling 1  0.328   0.328   1.138 0.308954
Canaritos     1  0.002   0.002   0.006 0.939691
Arbol         1  5.995   5.995  20.817 0.000813 ***
Residuals    11  3.168   0.288
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Primeras pruebas complementarias

Experimentos paralelos implementando segunda etapa de EXP Colab. :

+ RF: **50_7_200_200**
++ Canaritos: **0.75 y 2.0**
1.0 y 2.5

Factores fijos:

- ML y Deflacion
- Meses con pandemia (- marzo de 2020)
- Sin Julio de 2021 en el training
- Undersampling de 0.5
- K-folds en Cross Validation en 5
- Semillerío final con **50** semillas
- Sin regularización

Segundos resultados

17	5	0.3	0.5 y 0.5	25_6_200_200	sin regul	impar	47.031
18	5	0.3	0.5 y 0.5	50_7_200_200	sin regul	par	49.141
19	5	0.3	0.5 y 0.5	25_6_200_200	sin regul	polska	51.177
20	5	0.3	0.5 y 0.5	50_7_200_200	sin regul	impar 20	47.671
21	5	0.3	0.5 y 0.5	25_6_200_200	sin regul	polska	48.811
22	5	0.3	0.5 y 0.5	50_7_200_200	sin regul	par	47.463
23	5	0.3	0.75 y 2.0	25_6_200_200	sin regul	polska	48.229
24	5	0.5	0.75 y 2.0	50_7_200_200	sin regul	impar	47.353
25	5	0.5	1.0 y 2.5	50_7_200_200	sin regul	par	47.937
26	5	0.5	1.0 y 2.5	25_6_200_200	sin regul	polska	48.901
26 rank 5	5	0.5	1.0 y 2.5	25_6_200_200	sin regul	polska	
26 rank 12	5	0.5	1.0 y 2.5	25_6_200_200	sin regul	polska	menos de 49
26 rank 3	5	0.5	1.0 y 2.5	25_6_200_200	sin regul	polska	

Segundas pruebas complementarias

…“Volver a empezar ♫♫ “:

- Corrida de la mejor configuración del workflow del equipo A del EXP 06

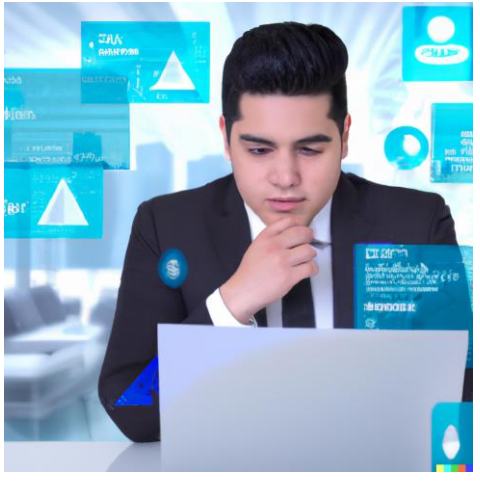
~~***Pico vs Valle en el Public y el Private***~~

- Experimentos seleccionando **otros rankings**
- Experimentos **hibridando semilleríos**

MAL... pero no TAN MAL!

imparcan1							50.149
imparcan5							52.325
imparcan10							53.087
CV al1	5	0.5	1.0 y 2.5	50_7_200_200	sin regul	al1	XX
CV al2	5	0.5	1.0 y 2.5	25_6_200_200	sin regul	al2	XX

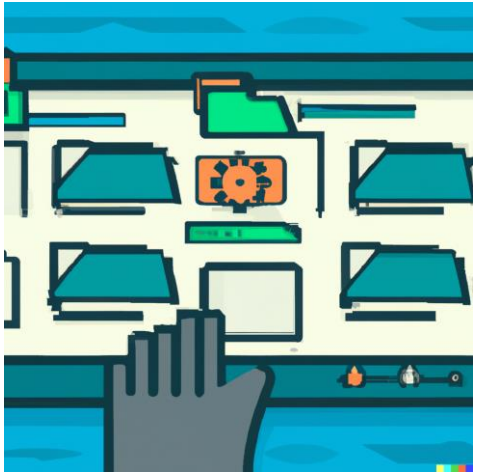
Experimento	Combinación de semillerios			Ganancia Públic	Max Public	Envíos
hibrid semillerios EXP1	18, 19, 20			50.361086		
hibrid semillerios EXP2	18, 19, 20, testigo			51.281066		
hibrid semillerios EXP3	19, testigo			51.455064		
hibrid semillerios EXP4	**11, 19, testigo**			**52.535048**	**53.07904**	**11000**
hibrid semillerios EXP5	11, 19, testigo, 23			52.109054		
hibrid semillerios EXP6	11, 18, 19, testigo, 23			52.139054		
hibrid semillerios EXP7	11, 18, 19, testigo, 21, 23			51.60706		
hibrid semillerios EXP8	**11, 19, testigo, 5, 12**			**52.501046**	**53.73902**	**10500**
hibrid semillerios EXP9	11, 19, testigo, 5			52.451044	53.32903	10500
hibrid semillerios EXP10	**11, 19, testigo, 12**			**52.987038**	**53.76902**	**11000**
hibrid semillerios EXP11	19, testigo, 12			52.619044		
hibrid semillerios EXP12	18, 20, 22, 24, 25	RF grandes		menos de 50		
hibrid semillerios EXP13	17, 19, 21, 23	RF medianos		menos de 51		
hibrid semillerios EXP14	1 a 8	RF más chicos		menos de 48		
hibrid semillerios EXP15	9 a 16	RF medianos		menos de 49		
hibrid semillerios EXP16	**11, 19, testigo, 12, 18**			**52.23705**	**53.75902**	**11000**
hibrid semillerío can 16				53.013238	53.11001	11000



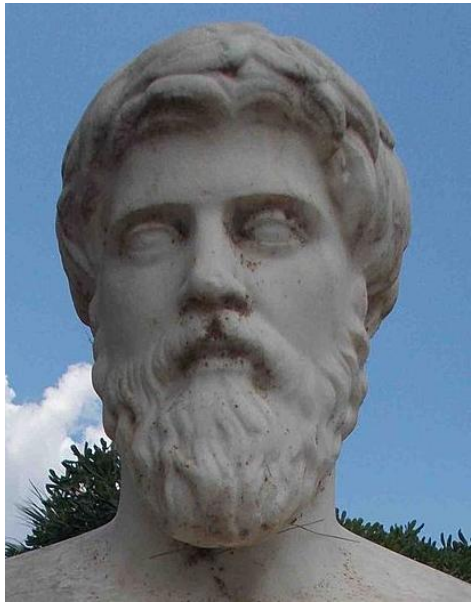
Conclusiones

- **Al menos se logró superar algo al Testigo**

- ***HIBRIDACIÓN SALVADORA***



Conclusiones



“Enseñar, más que llenar un recipiente es encender un fuego” (Plutarco)

GRACIAS!