



PROBLEMA #05

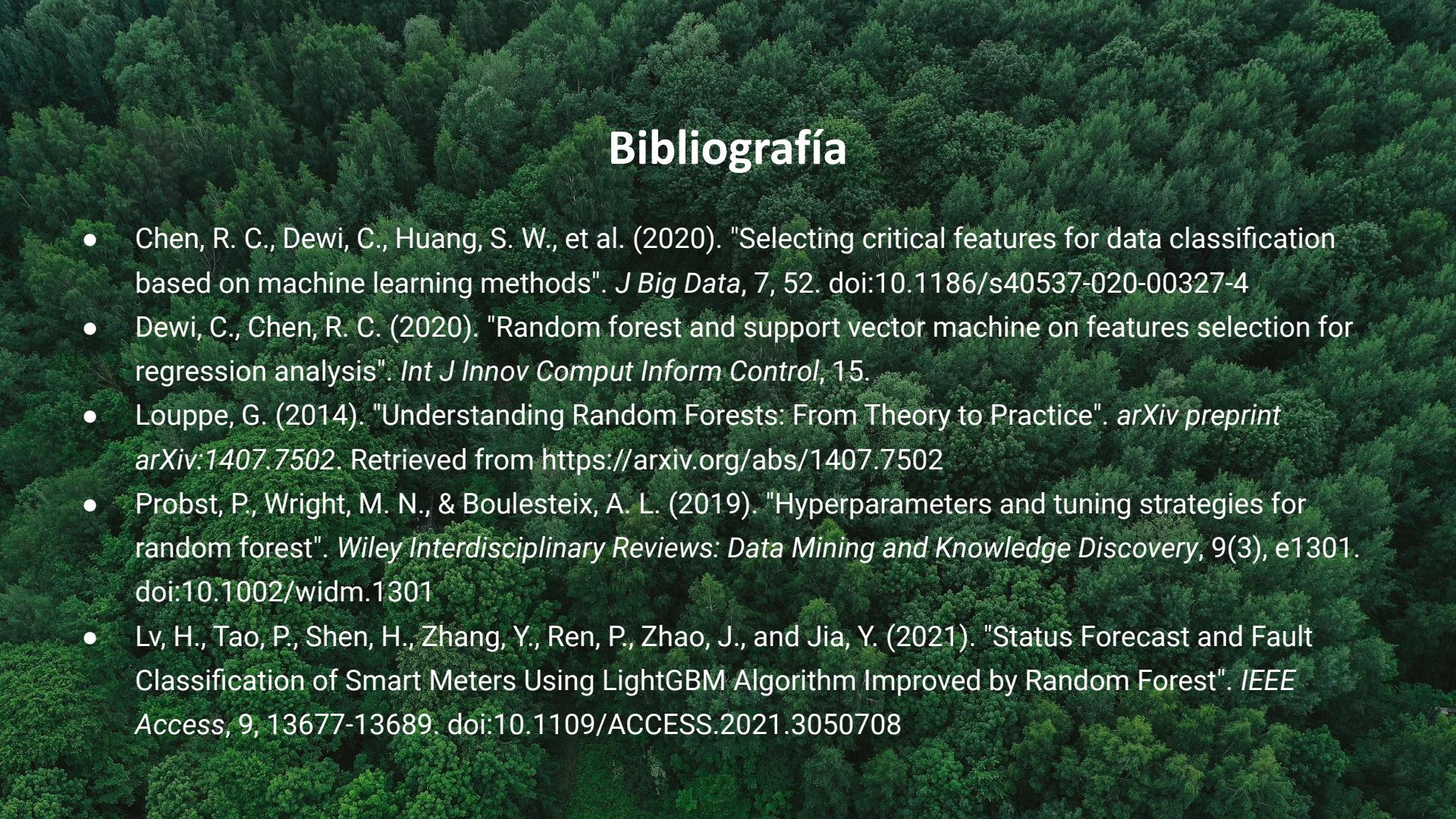
FE - FEATURE ENGINEERING INTRA MES

"VARIABLES DERIVADAS DEL RANDOM FOREST"

Grupo B: Agustina Ibarzábal

Hipótesis Experimental

Se postula que las variables generadas de manera automática por el método Random Forest serán significativas y contribuirán positivamente al aumento de las ganancias totales del modelo.

The background of the slide is a dark, grainy aerial photograph of a dense forest, showing a variety of green shades and some darker areas where trees are more closely packed.

Bibliografía

- Chen, R. C., Dewi, C., Huang, S. W., et al. (2020). "Selecting critical features for data classification based on machine learning methods". *J Big Data*, 7, 52. doi:10.1186/s40537-020-00327-4
- Dewi, C., Chen, R. C. (2020). "Random forest and support vector machine on features selection for regression analysis". *Int J Innov Comput Inform Control*, 15.
- Louppe, G. (2014). "Understanding Random Forests: From Theory to Practice". *arXiv preprint arXiv:1407.7502*. Retrieved from <https://arxiv.org/abs/1407.7502>
- Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). "Hyperparameters and tuning strategies for random forest". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. doi:10.1002/widm.1301
- Lv, H., Tao, P., Shen, H., Zhang, Y., Ren, P., Zhao, J., and Jia, Y. (2021). "Status Forecast and Fault Classification of Smart Meters Using LightGBM Algorithm Improved by Random Forest". *IEEE Access*, 9, 13677-13689. doi:10.1109/ACCESS.2021.3050708

Sesgos Cognitivos

Sesgos cognitivos respecto a los distintos valores de los hiperparametros:

- Mayor cantidad de arboles, mayor ganancia
- Mayor cantidad de hojas, mayor ganancia
- Menor feature fraction, mayor ganancia
- Mayor min_data_inleaf, mayor ganancia

Diseño Experimental

Se decide como diseño experimental realizar, en primer lugar, una corrida del workflow sin la función del random forest. Luego se procede a realizar distintas corridas alterando uno por uno los hiperparametros para observar cómo varía la ganancia y de acuerdo a la mejoría o no de la misma, probar nuevas combinaciones de hiperparametros.

Nº	num_iterations	num_leaves	min_data_in_leaves	feature_fraction_bynode
1	20	16	1000	0.2
2	-	-	-	-
3	20	16	1000	0.2
4	20	16	1000	1
5	20	16	1000	0.5
6	300	16	1000	0.5
7	20	100	1000	0.8
8	100	16	1000	0.8
9	20	16	3000	0.8
10	50	20	1500	0.8
11	50	100	1500	0.8
12	200	100	1500	1
13	300	200	2000	1
14	250	70	1500	1
15	300	16	1000	0.8
16	150	70	1200	0.8
17	150	60	500	1

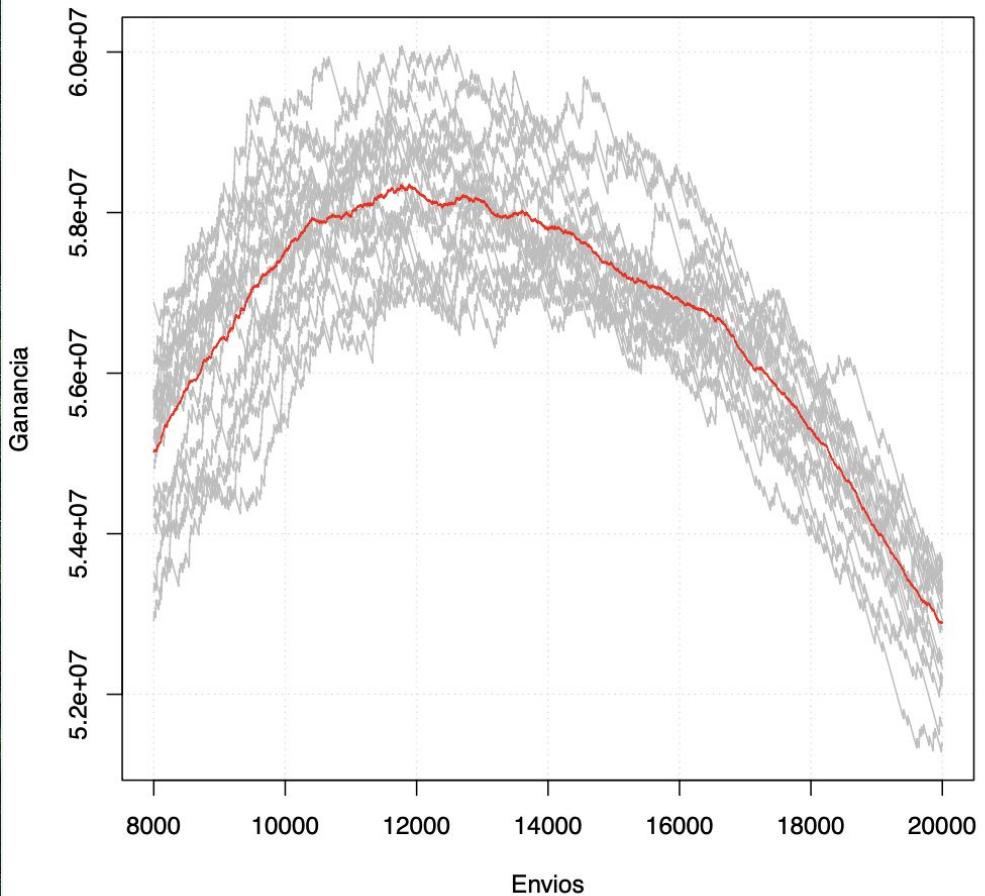
Limitaciones

- Google Cloud:
 - Los apagones de las maquinas
 - La memoria de las distintas instancias
- La falta de bibliografía respecto al uso de Random Forest en modelos LightGBM.
- El tiempo de corrida del código vs los días para realizar el experimento.

Resultados

Ganancias del modelo sin Random Forest
(experimento N°2)

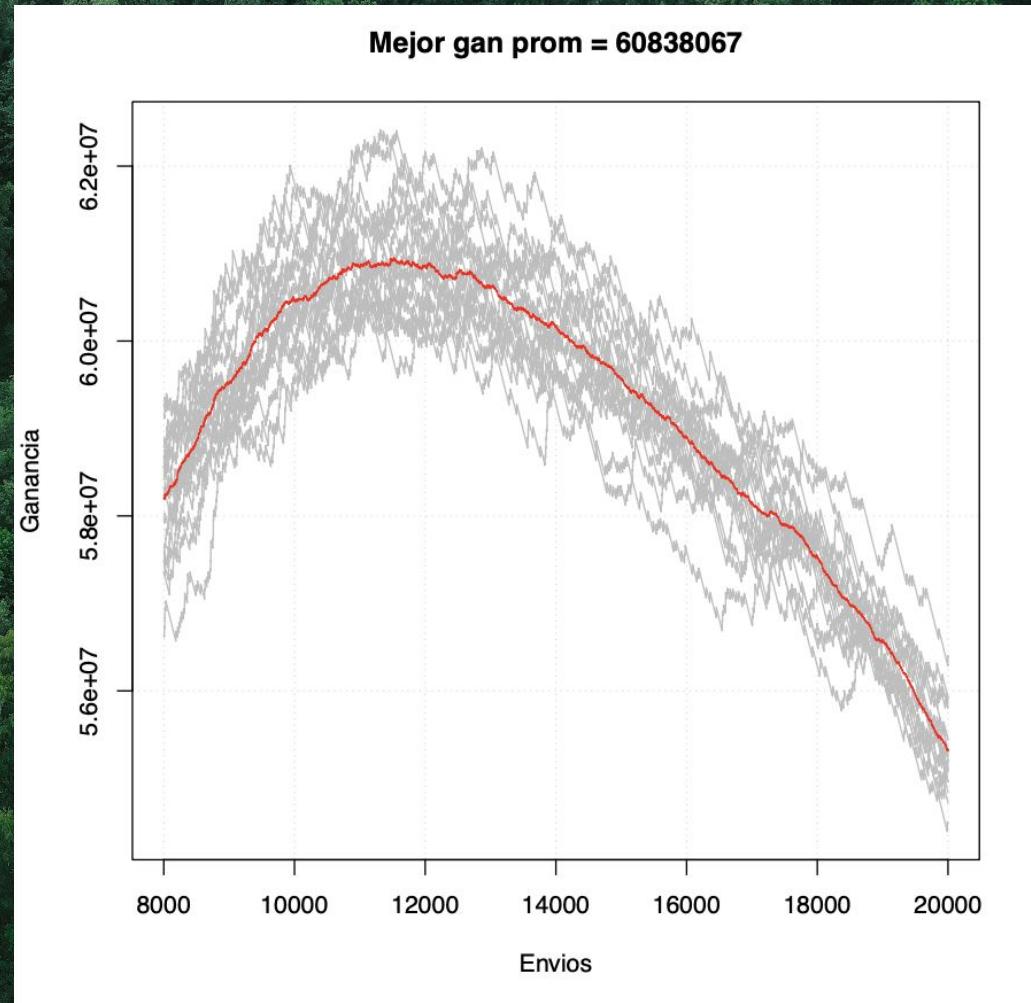
Mejor gan prom = 58179428



Resultados

Ganancias del modelo con Random Forest
(experimento N°6)

num_iterations: 300
num_leaves: 16
min_data_in_leaves: 1000
feature_fraction_bynode: 0.5



Resultados

Ranking	Control		Con RF	
	Feature	Gain	Feature	Gain
1	ctrx_quarter	0,061216419 rf_074_005		0,048699083
2	ctrx_quarter_normalizado	0,049701879 ctrx_quarter_normalizado		0,033548828
3	ctrx_quarter_lag1	0,41259493 rf_046_007		0,02166087
4	mcuentas_saldo_rank	0,027180804 rf_034_004		0,018069791
5	mprestamos_personales_rank	0,017559323 rf_203_000		0,016203341
6	cpayroll_trx	0,014998651 mcuentas_saldo_rank		0,013670415
7	mpasivos_margen_rank	0,014437454 rf_063_004		0,013170944
8	mpayroll_sobre_edad_rank	0,012969363 mprestamos_personales_rank		0,012371628
9	mtarjeta_visa_consumo_rank	0,010972314 mcaja_ahorro_rank		0,012336988
10	vm_status04	0,010447044 rf_038_006		0,012044303

Dentro de las 10 variables con mayor importancia, seis de ellas corresponden a variables generadas por Random Forest, siendo la primera una de ellas.

Resultados

Se realizó el test de Wilcoxon para comparar las medianas de ambas corridas.

Se obtuvo el siguiente resultado:

```
Wilcoxon signed rank test with continuity correction data:  
ganancias_01_035 and ganancias_01_062  
V = 270559302, p-value < 2.2e-16  
alternative hypothesis: true location shift is not equal to 0
```

Nº	num_iterations	num_leaves	min_data_in_leaves	feature_fraction_bynode	mejor ganancia promedio
1	20	16	1000	0.2	59582404
2	-	-	-	-	58179428
3	20	16	1000	0.2	59410838
4	20	16	1000	1	59794689
5	20	16	1000	0.5	59634485
6	300	16	1000	0.5	60838067
7	20	100	1000	0.8	60189284
8	100	16	1000	0.8	58571398
9	20	16	3000	0.8	58464490
10	50	20	1500	0.8	58672397
11	50	100	1500	0.8	56895401
12	200	100	1500	1	cannot allocate vector
13	300	200	2000	1	cannot allocate vector
14	250	70	1500	1	cannot allocate vector
15	300	16	1000	0.8	58877418
16	150	70	1200	0.8	56636676
17	150	60	500	1	60073364

Discusión de los resultados

En la mayoría de los casos, los experimentos en los que se incluyeron las variables generadas de manera automática por Random Forest, superaron las ganancias del experimento control. Los únicos que no cumplieron esta premisa fueron los Nº 6 y 11, por lo que también queda claro que esa combinación de hiperparámetros no es buena.

Se realizó el test de Wilcoxon que descarta la hipótesis de que las medianas entre ambos experimentos son iguales. Se puede decir que hay diferencia significativa entre la corrida control y la que tiene Random Forest.

Conclusiones

- La utilización de Random Forest mejora la ganancia del modelo.
- A mayor cantidad de iteraciones, mejores resultados.
- Si se utiliza mayor cantidad de iteraciones, no se debería aumentar mucho el número de hojas.
- Valores de feature fraction mayores a 0.5 mejoran la ganancia.

The background of the slide is a dark, grainy aerial photograph of a dense forest, showing a variety of green shades and tree patterns.

Recomendación Concreta

- num_trees entre 100 y 200
- num_leaves entre 20 y 40
- min_data_inleaves alrededor de 1000
- feature_fraction_bynode entre 0.5 y 1

Futuros Problemas y Experimentos

- En mi opinión se pueden seguir probando nuevas combinaciones.
- En concreto, me hubiese gustado probar valores menores de min_data_in_leaves ya que la corrida en la que le puse valor 500 me dio una buena ganancia, pero no llegué a hacer más. Probaría algún valor alrededor de 50-100.

Anexo

<https://github.com/agusibarzabal/ExperimentoColaborativo.git>