

# ¿Cómo elegir el mejor modelo?

UBA 2024

Mindstorms & Powerful Ideas

# Agenda

- Motivación
- Resultados Experimentales
- Test Estadístico para la comparación de Modelos Predictivos
- Pensando fuera de la caja

Focalice su energía en  
comprende profundamente  
temas de ciencias de datos,  
ya que aún no domina cuestiones básicas.

Luego vendrá la automatización.

# Motivación

# Motivación Futura

En un viaje al futuro de esta asignatura,  
a horas de finalizada Ultima Compencia Kaggle

reina la confusión, y en muchos casos, **enojo**.

# Motivación Public vs Private

Public   Private

This leaderboard is calculated with approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

#	Team	Members	Score	Entries	Last	Solution
4	Carla Campetella M		26.83817	35	3y	

Private

30	▼ 26	Carla Campetella M		22.12833	35	3y
----	------	--------------------	---	----------	----	----

# Motivación Private Leaderboard

39	▼ 21	Guillermo Rafael Epszteyn		21.74262	49	3y	
40	▼ 32	Pablo Acuña		21.70333	72	3y	
41	▼ 24	Belen Bernatene		21.70155	86	3y	
42	▼ 8	Alhussain Maalla		21.65155	47	3y	
43	▼ 22	Chapulin Colorado	 	21.63726	97	3y	
44	▼ 31	Yamila Rodriguez		21.63369	97	3y	
45	▼ 7	Florencia Varise		21.60155	175	3y	

# Motivación Private Leaderboard

53	▼ 22	Cristian De Blasis		21.13013	329	3y	
54	▲ 5	Pablo Marenco		21.06227	21	3y	
55	▼ 40	ricve2001@gmail.com		21.01227	90	3y	

# Motivación Private Leaderboard

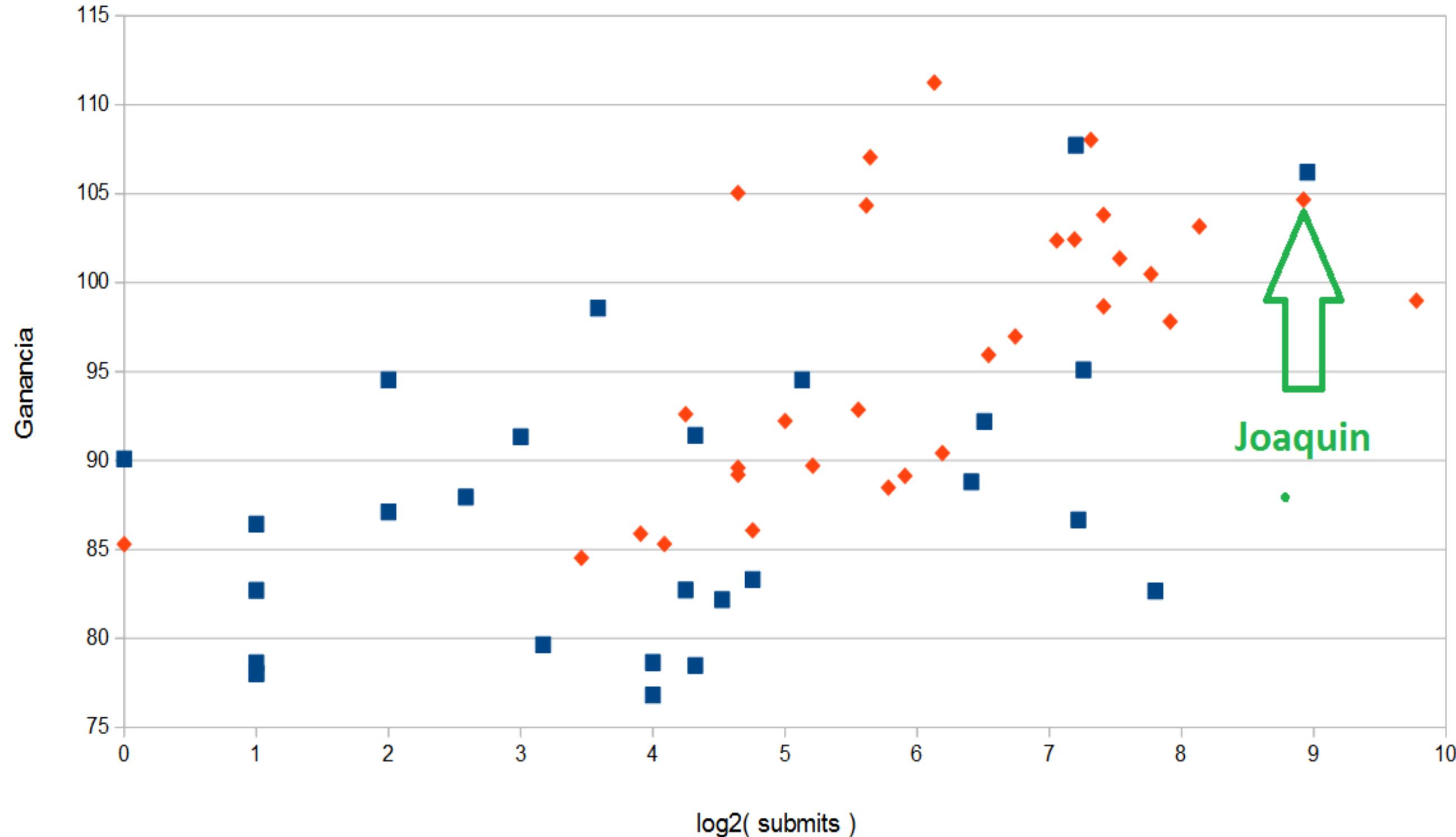
39	▼ 21	Guillermo Rafael Epszteyn		21.74262	49	3y	
40	▼ 32	Pablo Acuña		21.70333	72	3y	
41	▼ 24	Belen Bernatene		21.70155	86	3y	
42	▼ 8	Alhussain Maalla		21.65155	47	3y	
43	▼ 22	Chapulin Colorado	 	21.63726	97	3y	
44	▼ 31	Yamila Rodriguez		21.63369	97	3y	
45	▼ 7	Florencia Varise		21.60155	175	3y	

# Motivación comentarios

"No comparto que el mayor porcentaje de la nota final sea el Private Leaderboard, algo azaroso que nunca vemos, no se relaciona en el Public y no hay forma cierta de optimizarlo"

"Es como si 10 o 20 puntos porcentuales de tu nota dependan exclusivamente de la puntuación de tirar un dado, no hay ningún mérito en un elemento azaroso que solo podés ver una vez cerrada la nota. Sería mucho más justo considerar TODOS los submits de un alumno o que puntaje público = privado."

## Competencia UNO Ganancia vs Submits



# Motivación

comentarios

"Nunca termine de entender cuál fue mi mejor modelo y que podría haber hecho para que sea mejor."

"....da la sensación de que en realidad estuvimos todo el tiempo en un **casino** jugando a la **ruleta**."

"de la nada un modelo al que le **apostas** todo no sirve para nada."

## Motivación comentarios

Un alumno de mitad de tabla propone

"Capaz se me ocurre que los primeros puestos nos expliquen directo que hicieron , para aprender."

Un integrante del equipo ganador, responde

"Si supiera, te lo diría."

El otro integrante del equipo ganador

"Del Público al Privado subimos 42 posiciones, me cuesta entender por que se produjo tanta diferencia"

## Motivación comentarios

- "a mi me gustaría entender todas las situaciones, no solo los primeros puestos:
  - Los q estaban en el primer puesto en el público pero cayeron muchos puestos en el privado
  - Los que estaban abajo en el público pero en el privado subieron muchos puestos"
- "realmente me cuesta expresar lo mucho que me he esforzado, y la poca esperanza que tengo de encontrar una solución verdadera (y no depender únicamente del azar dentro de un margen de error)"

## Motivación comentarios

"... no puedo conectar todo el análisis hecho con los resultados de Kaggle."

"Me voy a quedar con la imagen del Publico!!! (Ojos que no ven.....)"

"Yo tuve una gran frustración al ver en el privado la verdad, sigo pensando hoy qué métodos podría haber utilizado para que las señales, por las que elegí el que elegí, me haya dado el correcto."

## Motivación comentarios

"a pesar que las ultimas dos semanas le dediqué todo mi tiempo libre, me decepcionaron mis resultados en el Private. Pero lo tomo como un aprendizaje...

Lo que me desvela es : en modelos de mi trabajo ¿cómo reducir la variabilidad de la predicción?, ya que seguramente me está sucediendo y recién esta materia me abrió los ojos.  
¿En cuántos modelos habré tenido mala suerte por el azar?"

# Motivación

nombres de equipos Kaggle

- Monos que apretan palancas
- Team **Suerte** y Overfitting



# Motivación

Public   Private

The private leaderboard is calculated with approximately 70% of the test data.  
This competition has completed. This leaderboard reflects the final standings.

#	△	Team	Members	Score	Entries	Last	Solution
1	▲ 38	Monos que apretan palancas	 	23.83903	209	3y	
2	▲ 8	Team AD&SJ	 	23.71760	307	3y	
3	▲ 42	Team Suerte y Overfitting	 	23.44974	302	3y	
4	▲ 37	gerbeldo		23.06046	40	3y	
5	▲ 17	Agustina Stekolschik		23.05153	96	3y	

Para la gran mayoría (que no conocíamos algo mejor) la única forma de evaluar si lo que probaba estaba funcionando era EL PUBLICO de Kaggle trabajando siempre con la misma semilla, y teníamos 5 intentos cada 24hs.

No se hizo especial hincapié al comienzo en que implicaba el score de Kaggle, las distinciones entre público y privado, etc. Las sucesivas explicaciones a medida que avanzaba la desesperación del conjunto, no hacían mas que embarrar la cancha.

¿Cuánto puede variar la ganancia de un modelo?

¿Qué relación hay entre la ganancias medidas en 5-fold cross validation, Public y Private?

¿ Si un modelo M1 da más ganancia que M2 en 5-fold cross validation, también es mejor en el Public Leaderboard? ¿ y en el Private?

Para entender lo que está sucediendo  
se realizaron dos experimentos,  
al final se compararán dos modelos

# Experimento 1

## un modelo muy simple

# Experimento 1 Objetivo

Objetivo: analizar la variabilidad de un modelo *fijo*  
a partir de muy buenos parámetros  
que entrena solo en noviembre-2020

Finalmente, se observa el comportamiento de **volver a generar**  
**cada vez** el modelo con distintas semillas en:

- 5-fold cross validation
- Public Leaderboard
- Private Leaderboard

# Experimento 1 jugando con la semilla

¿Cuál es la variabilidad de las ganancias de LightGBM si se entrena en el mismo dataset, se dejan los hiperparámetros fijos, pero se cambia únicamente la semilla ( que sería lo mismo que reordenar al azar las columnas del dataset) ?

o sea, ¿Cuál es la variabilidad inherente de un modelo, generado en este caso con LightGBM ?

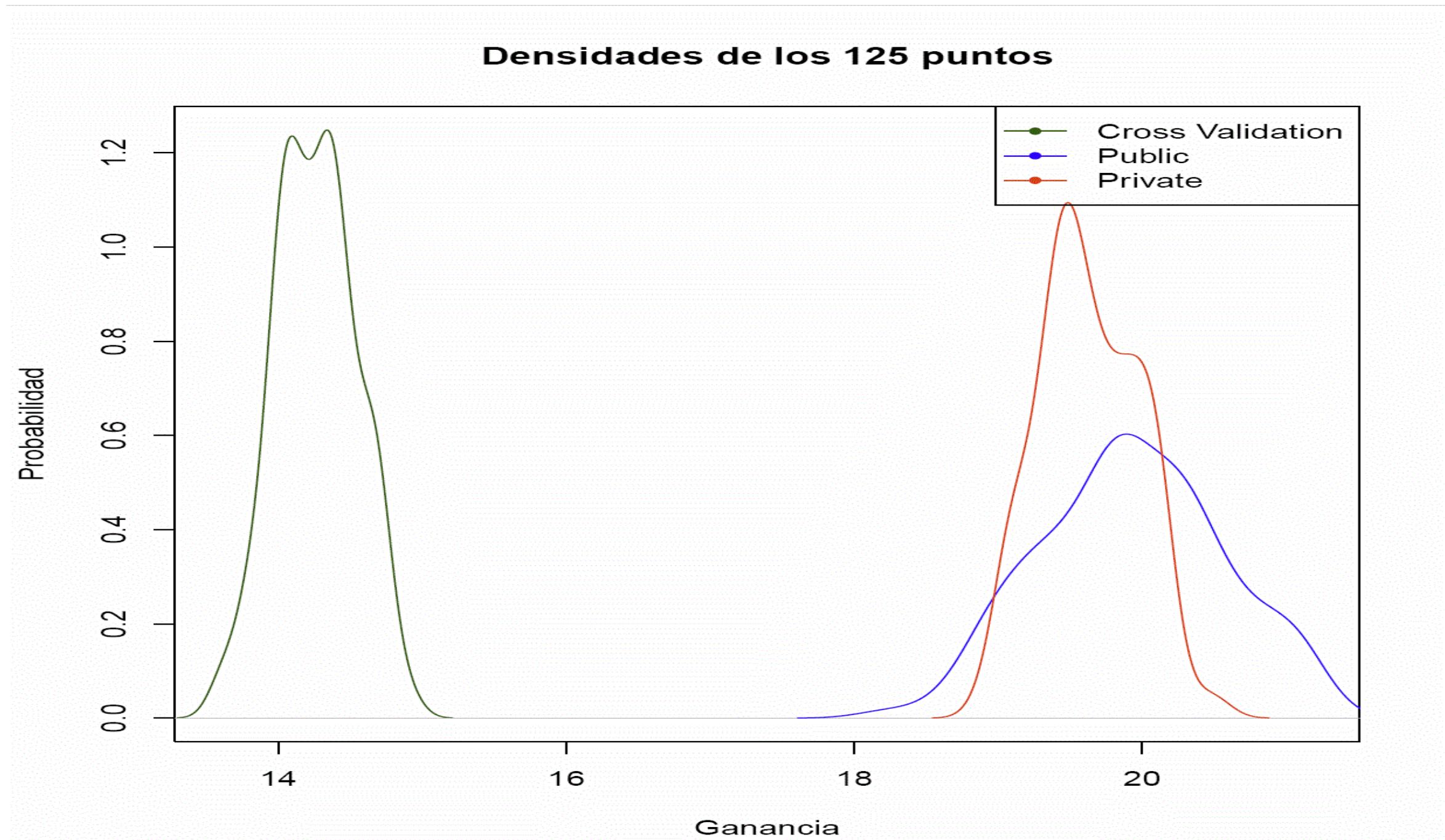
# Experimento 1 resultados

Cambiando las semillas las corridas jamás dan la misma ganancia ni en 5-fold cross validation, ni en el Public ni en el Private Leaderboard. Se graficará la función de distribución de probabilidad de esa variable aleatoria (la ganancia).

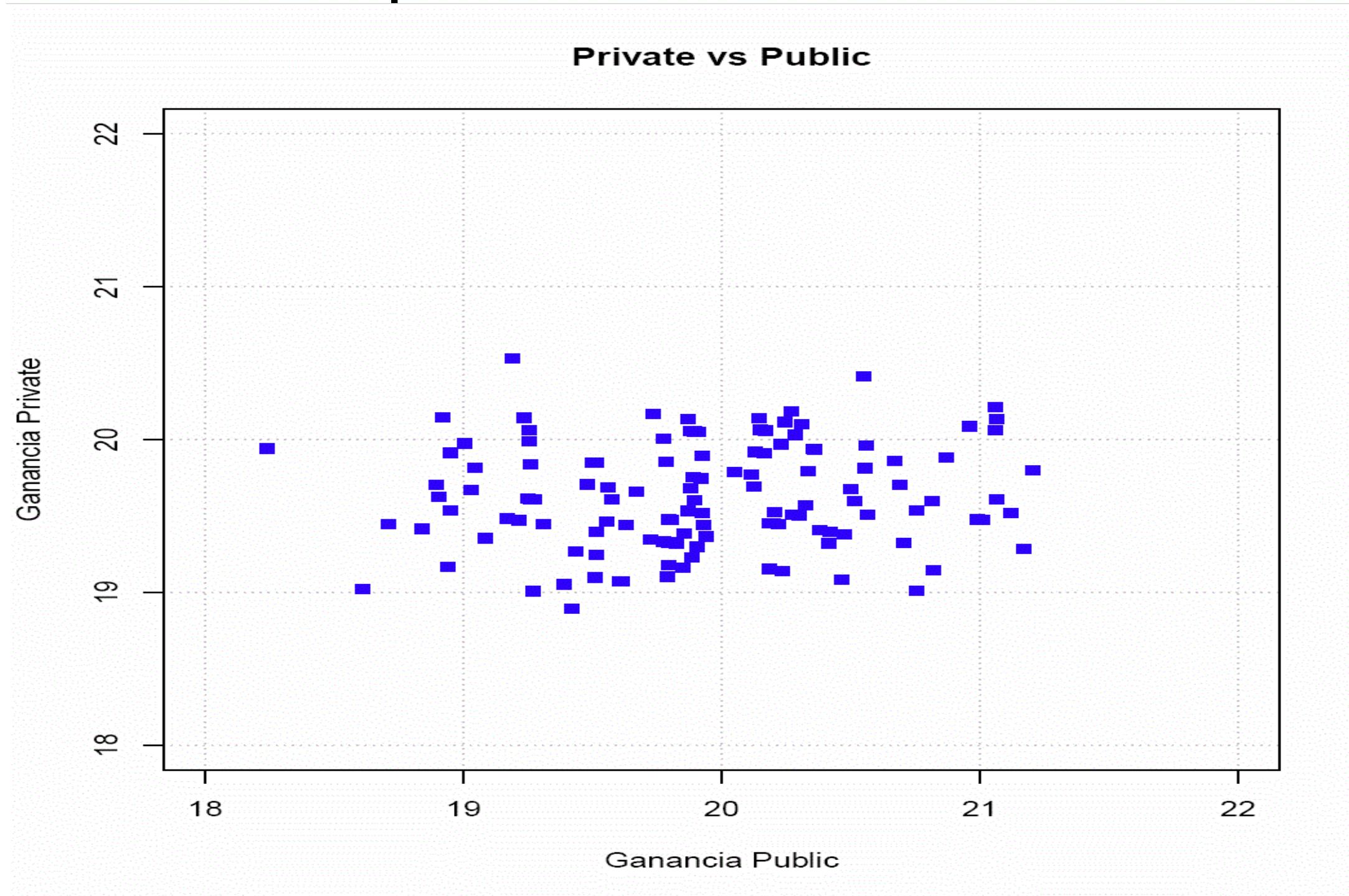
# Experimento 1 resultados

Ganancia	mean	sd
Cross Validation	14.3	0.28
Public	19.9	0.63
Private	19.6	0.34

# Experimento 1 resultados



# Experimento 1 aleatoriedad



# Overfitting the Leaderboard in Ernst & Young Data Science Competition 2019

And subsequently losing 8000 USD + a ticket to New York.



Ilham Firdausi Putra · [Follow](#)

Published in HMIF ITB Tech · 7 min read · Jul 15, 2019



587



Rank	Team	Score
1	christianwbsn, ilhamfp31	0.90101
9	christianwbsn, ilhamfp31	0.88815

#HMIFTECH

## What Went Wrong?

It was the cornerstone of an inept data scientist, stemming from a complete lack of experience. The main ingredient of a classic shake-up between public and private leaderboard score:

We did not trust our cross-validation score

— and the result was catastrophic. We had overfitted the public leaderboard.

It's easy to say that you believe in something until it doesn't align with what you see. In our case, we believed in our cross-validation score as it was in line with our leaderboard score — until it wasn't. At that point, any experienced data scientist competition enthusiast or Kaggle would know that it was perfectly reasonable and possible. Our lack of experience caused us to falter and eventually lose the position.

The public leaderboard was calculated with only a fraction of the total test data, which after the competition we know was around 1/3 public and 2/3



# Greg Park

[traitlab](#) · [research](#) · [software](#) · [writing](#) · [contact](#)

---

## The dangers of overfitting: a Kaggle postmortem

July 06, 2012

	 8	Random Forest Benchmark	0.86141		
45	 8	dickoa	0.86141	1	Tue, 22 May 2012 12:07:36
45	 8	Rohit	0.86141	2	Fri, 25 May 2012 21:00:14
45	 8	squawkboxed	0.86141	1	Fri, 08 Jun 2012 14:57:28
45	new	BLetson	0.86141	3	Fri, 29 Jun 2012 14:49:38
50	 9	testing	0.86135	4	Sat, 16 Jun 2012 05:18:44 (-26.1h)
51	 9	schappi	0.86130	7	Sat, 16 Jun 2012 12:53:13
52	 8	Greg Park	0.86116	42	Fri, 29 Jun 2012 01:08:38 (-14.1d)
53	 8	Glen	0.86111	35	Tue, 05 Jun 2012 23:44:06 (-3.3d)

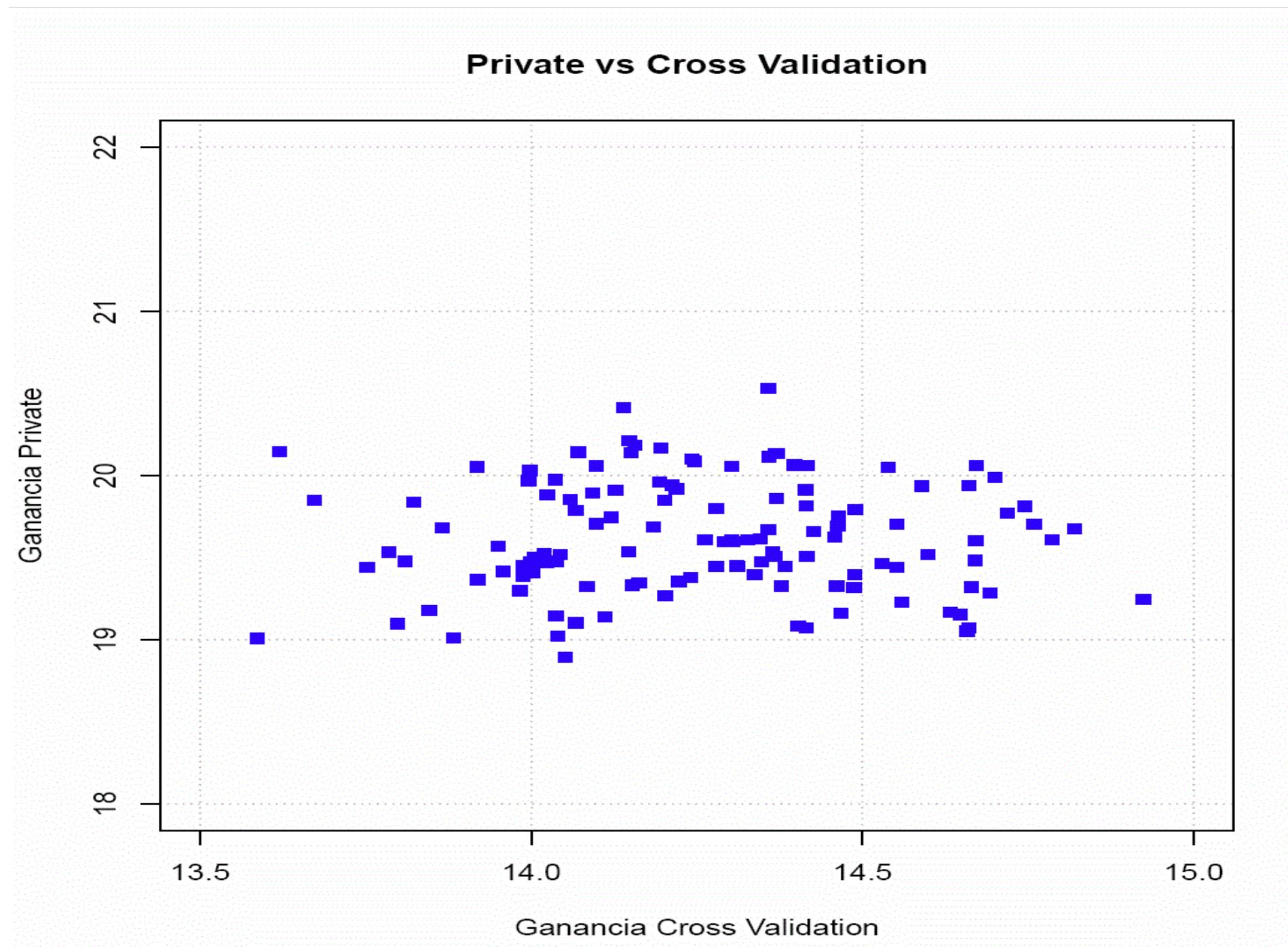
*My place on the private leaderboard. I dropped from 2nd place on the public leaderboard to 52nd on the private leaderboard. Notice I placed below the random forest benchmark!*

# Lessons learned

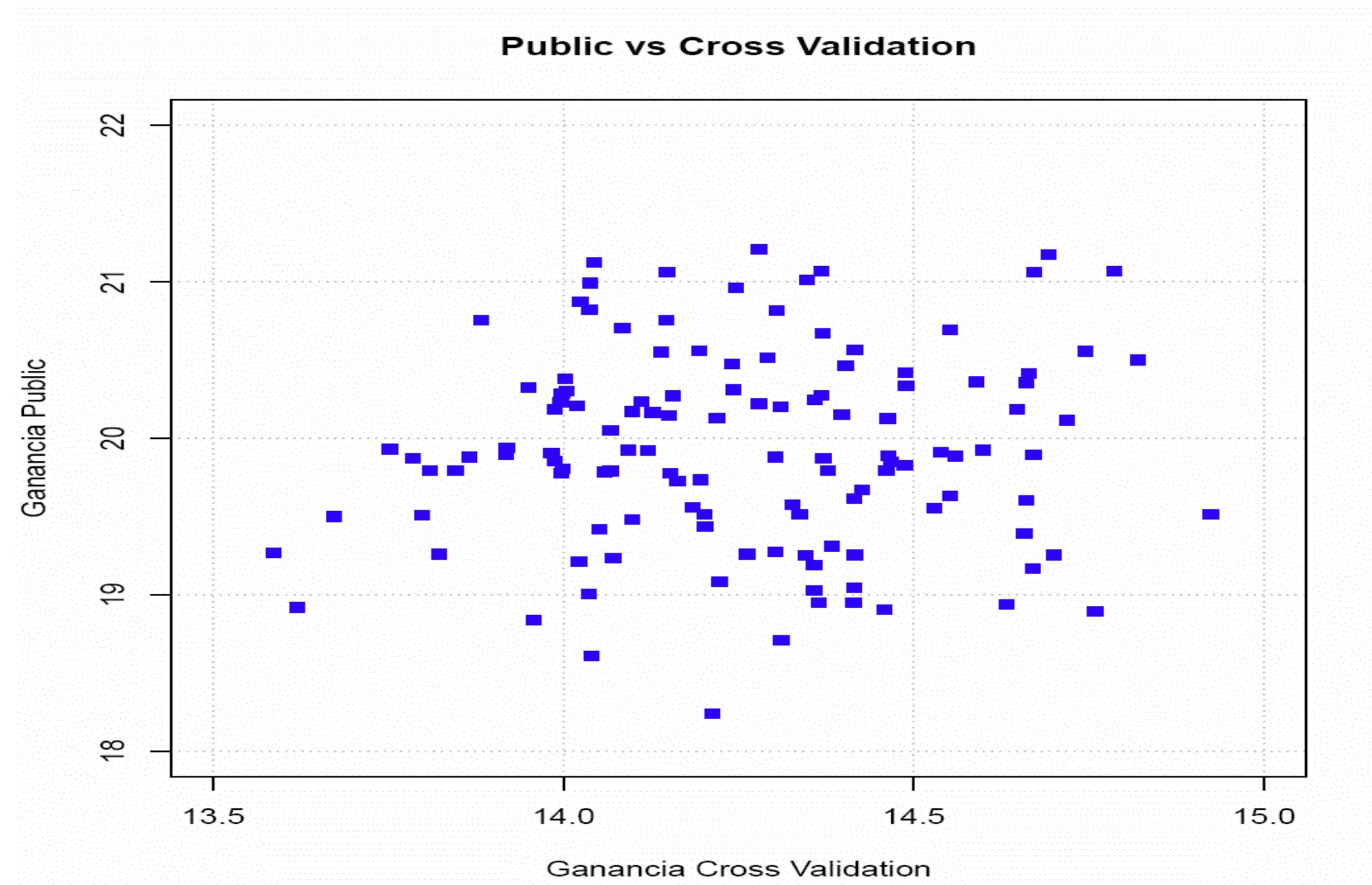
In the end, my slow climb up the leaderboard was due mostly to luck. I chose my five final submissions based on cross-validation estimates, which turned out to be a poor predictor of true score. Ultimately, I did not include my best submissions in the final five, which would have brought me up to 33rd place – not all that much better than 52nd. All said, this was my most educational Kaggle contest yet. Here are some things I'll take into the next contest:

- It is easier to overfit the public leaderboard than previously thought. Be more selective with submissions.
- On a related note, perform cross-validation the right way: include all training (feature selection, preprocessing, etc.) in each fold.
- Try to ignore the public leaderboard, even when it is telling you nice things about yourself.

# Experimento 1 aleatoriedad



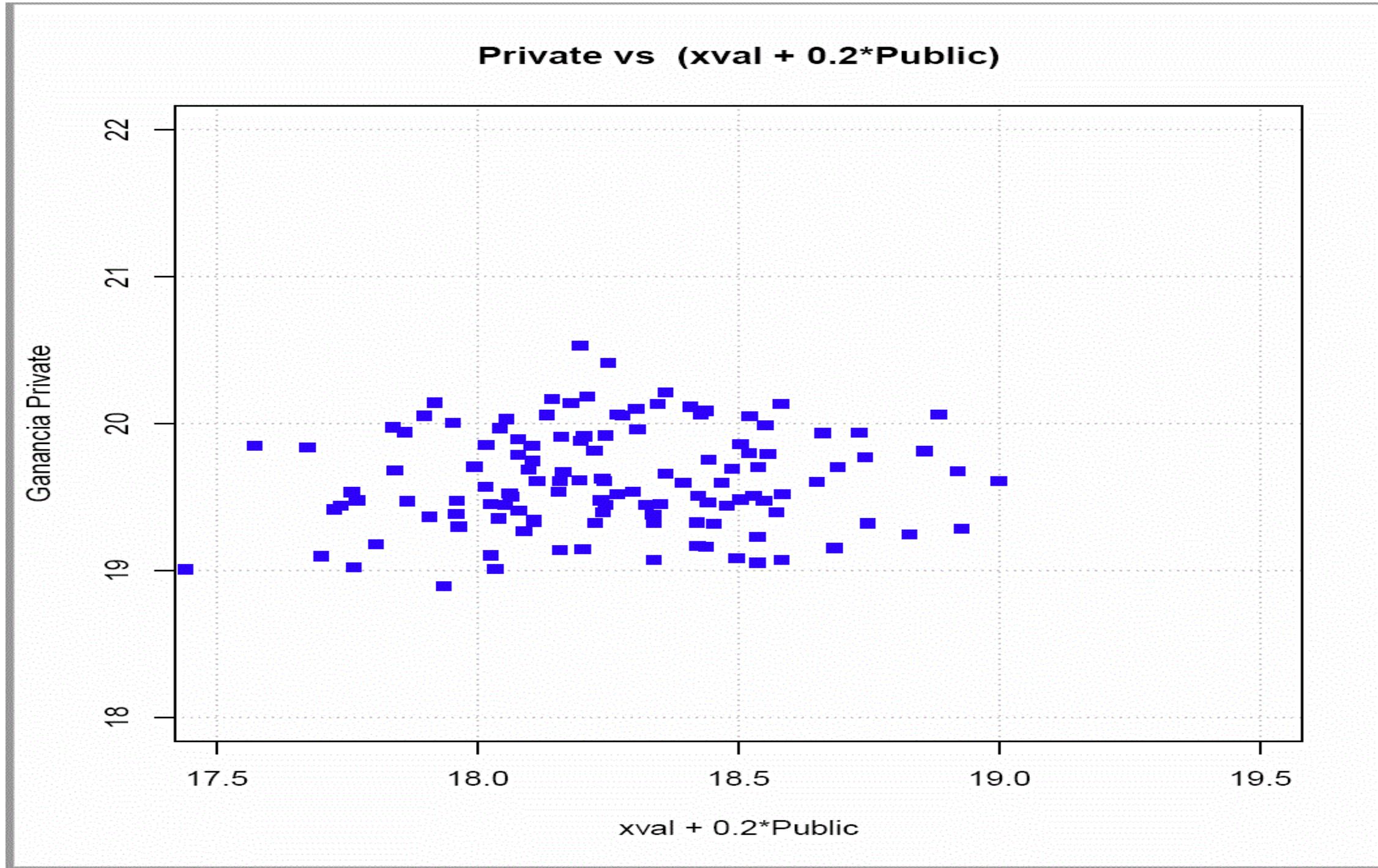
# Experimento 1 aleatoriedad



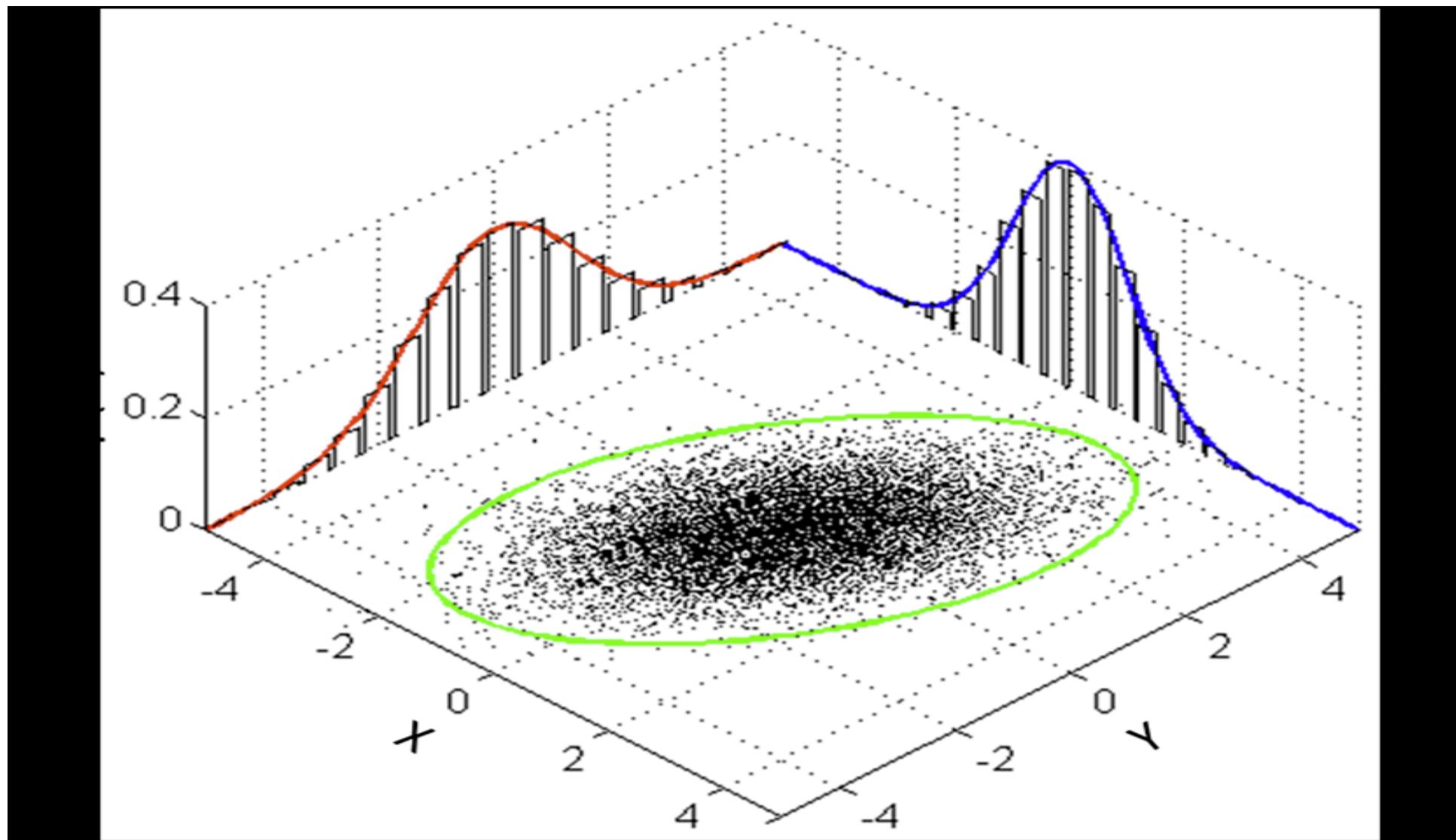
# Experimento 1 disgresión

¿ y si calculo en promedio (ponderado quizas) de las ganancias en los datasets de testing y Public Leaderboard, podré predecir mejor el Private ?

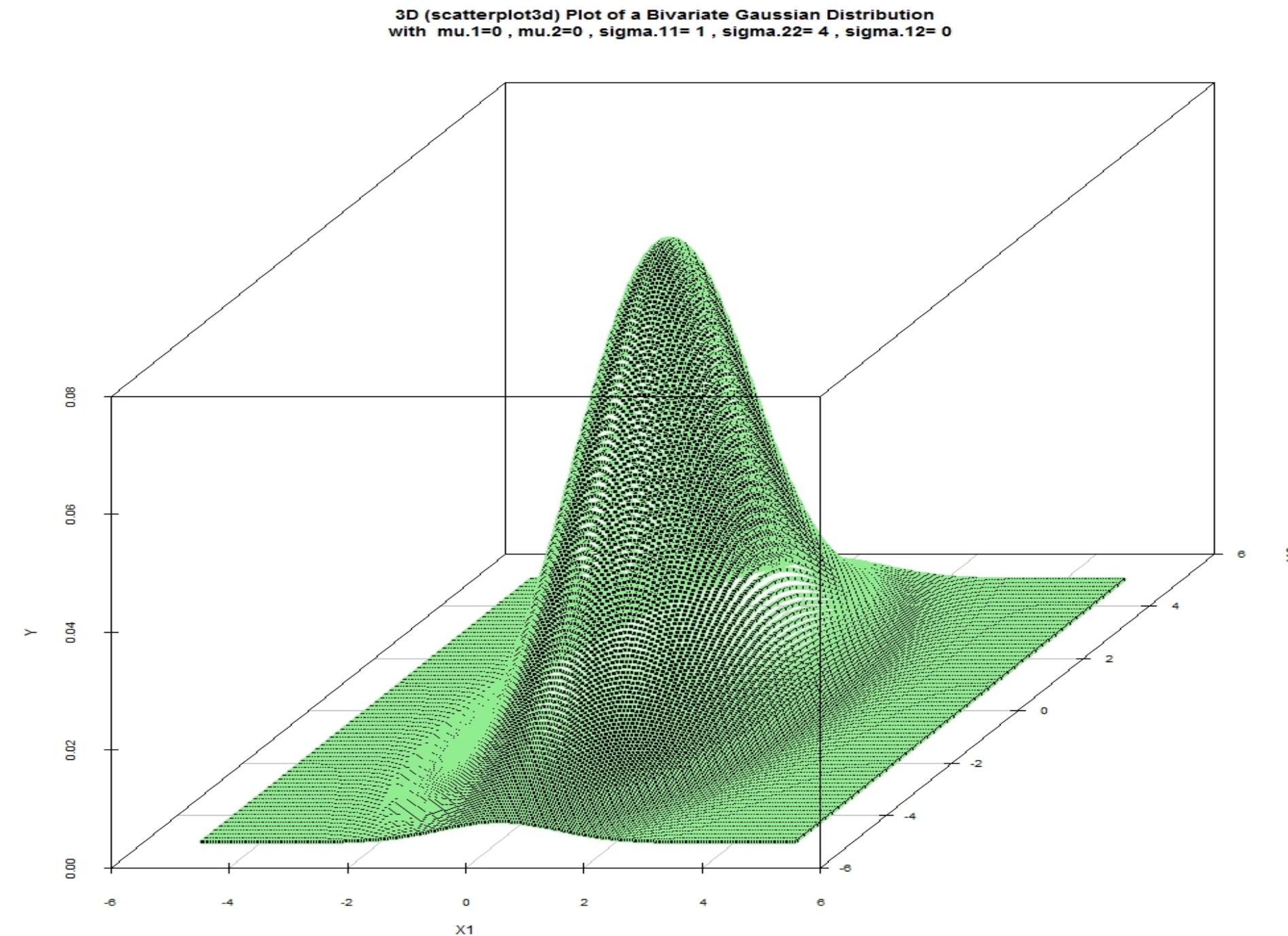
# Experimento 1 aleatoriedad



# Experimento 1 aleatoriedad



# Experimento 1 aleatoriedad



# Experimento 1 conclusión

- Los resultados 5-fold cross validation, Public y Private poseen una distribución *quasi* normal, y en caso que solo cambie la semilla son independientes entre si.
- No se puede saber si se va a estar por encima o por debajo de la media en los datos del futuro, por más que en el dataset que conozco si lo esté.

¿Cómo comparo dos modelos distintos, que fueron generados con datasets e hiperparámetros diferentes?

Comparar los dos motivacionales !

# Experimento 2

## Experimento 2 Objetivo

Objetivo: analizar la variabilidad de un modelo *fijo* que utiliza diez meses [202001, 202011] – 202006 dataset con lag1 y delta1

Finalmente, se observa el comportamiento de **regenerar** el modelo con distintas semillas en:

- 5-fold cross validation
- Public Leaderboard
- Private Leaderboard

## Experimento 2

Al dataset original ahora se le agregan los **lags y delta lag de orden 1**, además de corregir las variables *rotas*. Se buscan los hiperparámetros óptimos del LightGBM con una Optimización Bayesiana, train=[202001,202010] test=[202011]

Finalmente, se observa el comportamiento de **regenerar** el modelo con distintas semillas en:

- Testing , [202011]
- Public Leaderboard
- Private Leaderboard

## Experimento 2 dataset

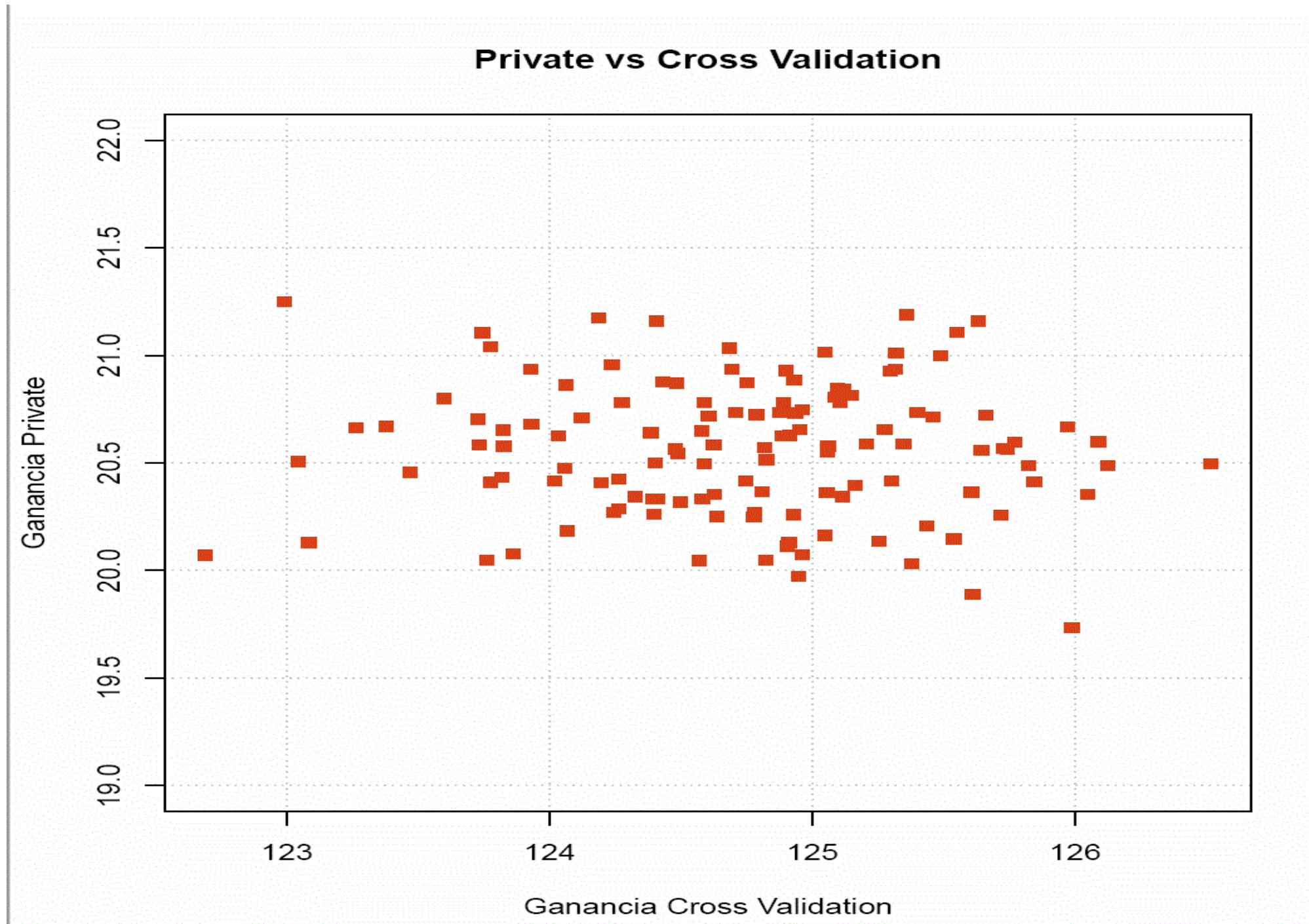
Para una variable, el lag de orden 1, `lag1` es el valor de esa variable el mes anterior. Si el mes anterior el registro no está en la base de datos, se asigna NA.

El `delta1` para una variable es el valor en el mes actual de la variable menos su valor el mes anterior.

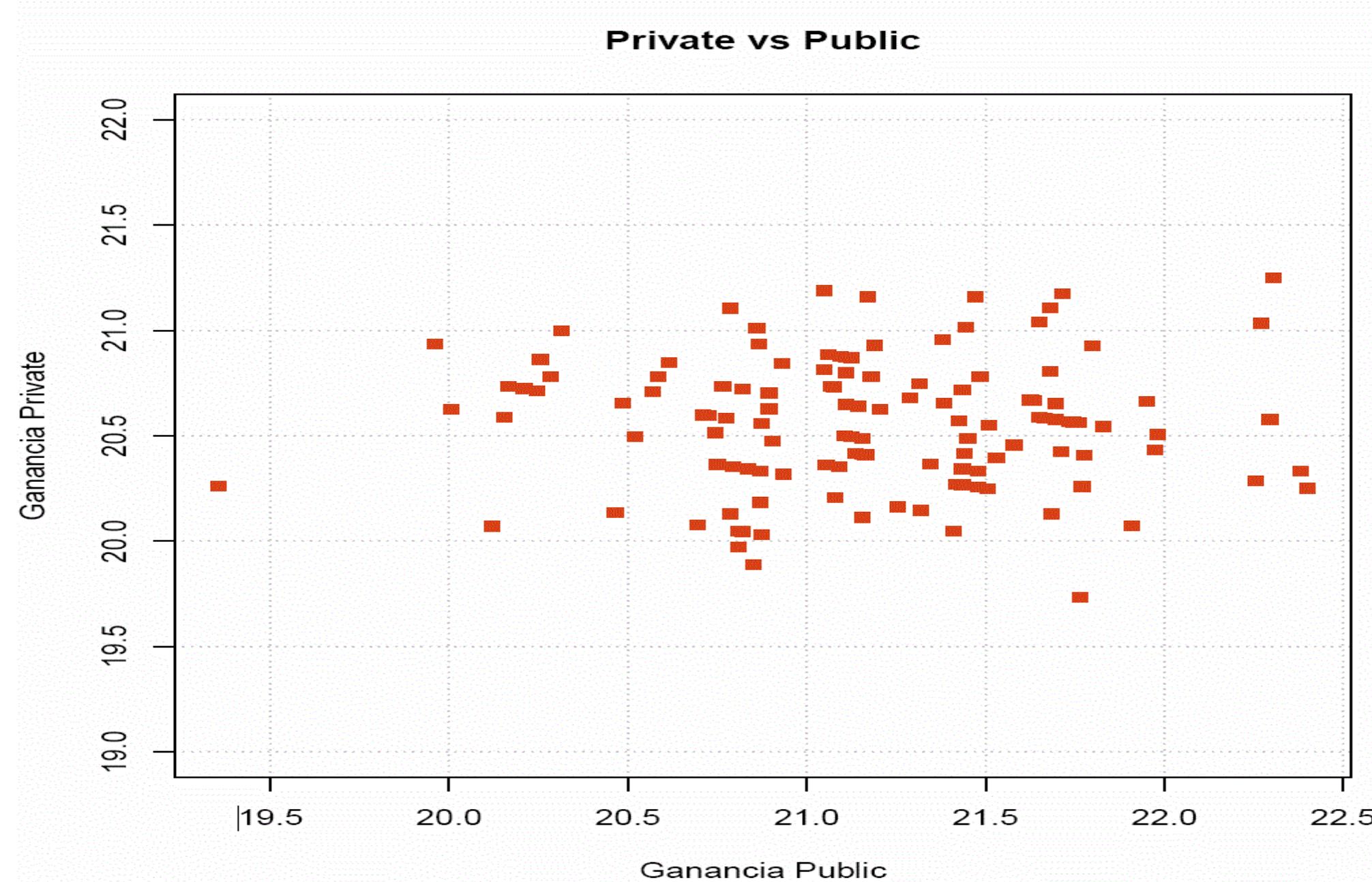
# Experimento 2 vs 1

Métrica	Variables Originales	Lag 1 + Delta1 10 meses
Cross Validation	14.3	124.7
Public	19.9	21.2
Private	19.6	20.6

# Experimento 2 aleatoriedad

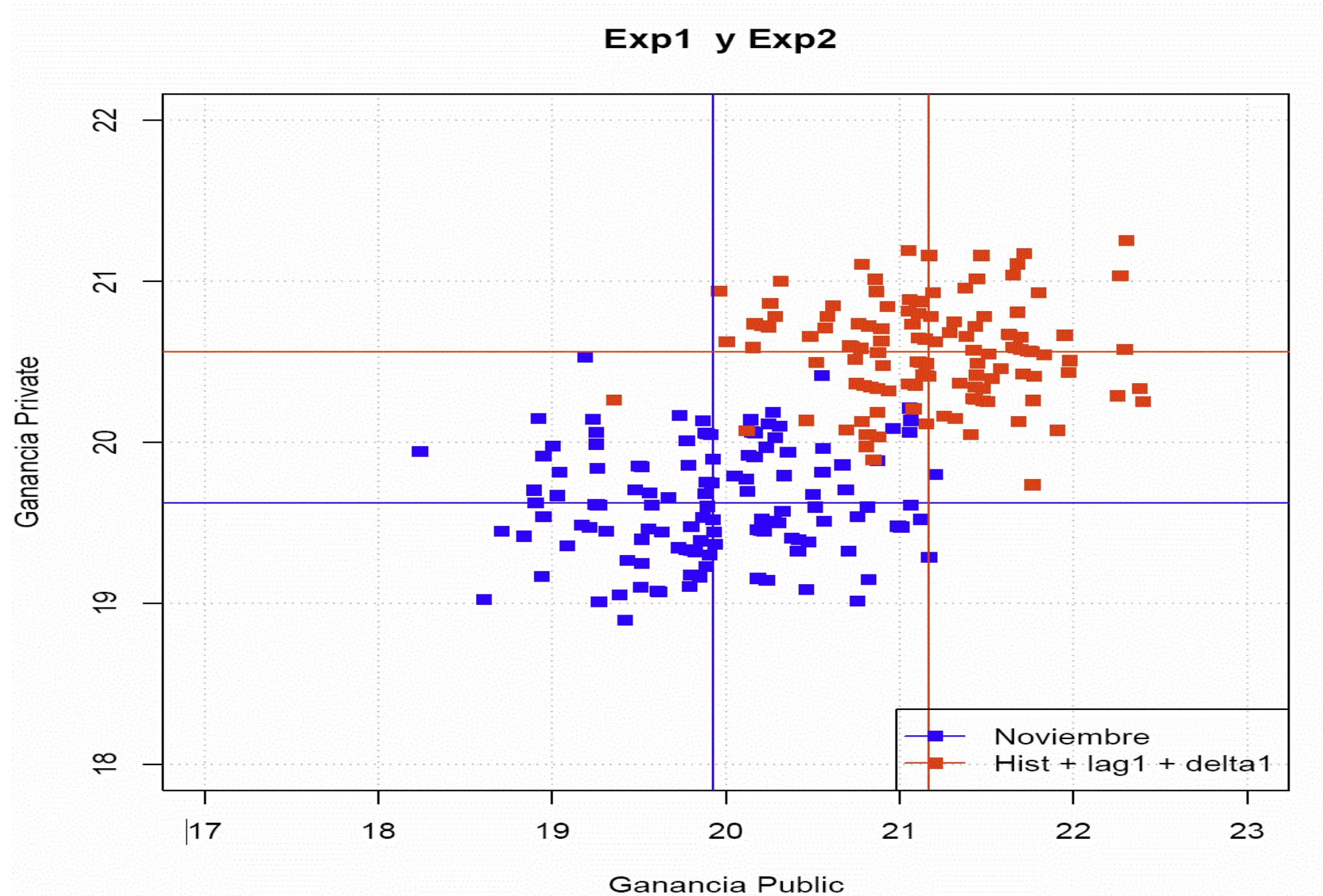


# Experimento 2 aleatoriedad



Finalmente, la comparación  
entre los dos experimentos

# Experimento 2 vs 1



## Conclusion General:

La comparación entre dos modelos predictivos M1 y M2 viene acompañada de una probabilidad.

Siempre se debe decir por ejemplo

`metrica(M2) > metrica(M1)`

con una probabilidad `p`

en el caso que `p` sea cercana a 0.5 hace falta un mayor número de observaciones para determinar el sentido de la desigualdad.

# Comparación estadística

Demsar, Janez. *Statistical Comparisons of Classifiers over Multiple Data Sets*, Journal of Machine Learning Research 7 (2006) 1–30, 2006

Wilcoxon signed rank test

en lenguaje R

```
wilcox.test( ganancias1, ganancias2, paired=TRUE)
```

ver script

qty	modelo1 campos orig Gan Public	modelo2 lag1 + delta1 GanPublic	Wilcoxon p-value
1	20.55	21.95	1.000

`wilcox.test( c(20.55), c(21.95), paired=TRUE) → 1.0`

Como **1.0** es mayor que 0.05, hacen falta más semillas

qty	modelo1 campos orig Gan Public	modelo2 lag1 + delta1 GanPublic	Wilcoxon p-value
1	20.55	21.95	1.000
2	19.78	20.29	0.667

```
wilcox.test(  
  c(20.55, 19.78),  
  c(21.95, 20.29),  
  paired=TRUE)
```

→ 0.667

Como 0.667 es mayor que 0.05, hacen falta más semillas

qty	modelo1 campos orig Gan Public	modelo2 lag1 + delta1 GanPublic	Wilcoxon p-value
1	20.55	21.95	1.000
2	19.78	20.29	0.667
3	20.76	20.87	0.400

```
wilcox.test(
  c(20.55, 19.78, 20.76),
  c(21.95, 20.29, 20.87),
  paired=TRUE)
```

→ 0.400

Como 0.400 es mayor que 0.05, hacen falta más semillas

qty	modelo1 campos orig Gan Public	modelo2 lag1 + delta1 GanPublic	Wilcoxon p-value
1	20.55	21.95	1.000
2	19.78	20.29	0.667
3	20.76	20.87	0.400
4	18.95	21.05	0.114

```
wilcox.test(
  c(20.55, 19.78, 20.76, 18.95),
  c(21.95, 20.29, 20.87, 21.05),
  paired=TRUE)
```

→ 0.114

Como 0.114 es mayor que 0.05, hacen falta más semillas

qty	modelo1 campos orig Gan Public	modelo2 lag1 + delta1 GanPublic	Wilcoxon p-value
1	20.55	21.95	1.000
2	19.78	20.29	0.667
3	20.76	20.87	0.400
4	18.95	21.05	0.114
5	19.62	21.68	0.032

```
wilcox.test(
  c(20.55, 19.78, 20.76, 18.95, 19.62),
  c(21.95, 20.29, 20.87, 21.05, 21.68),
  paired=TRUE)
```

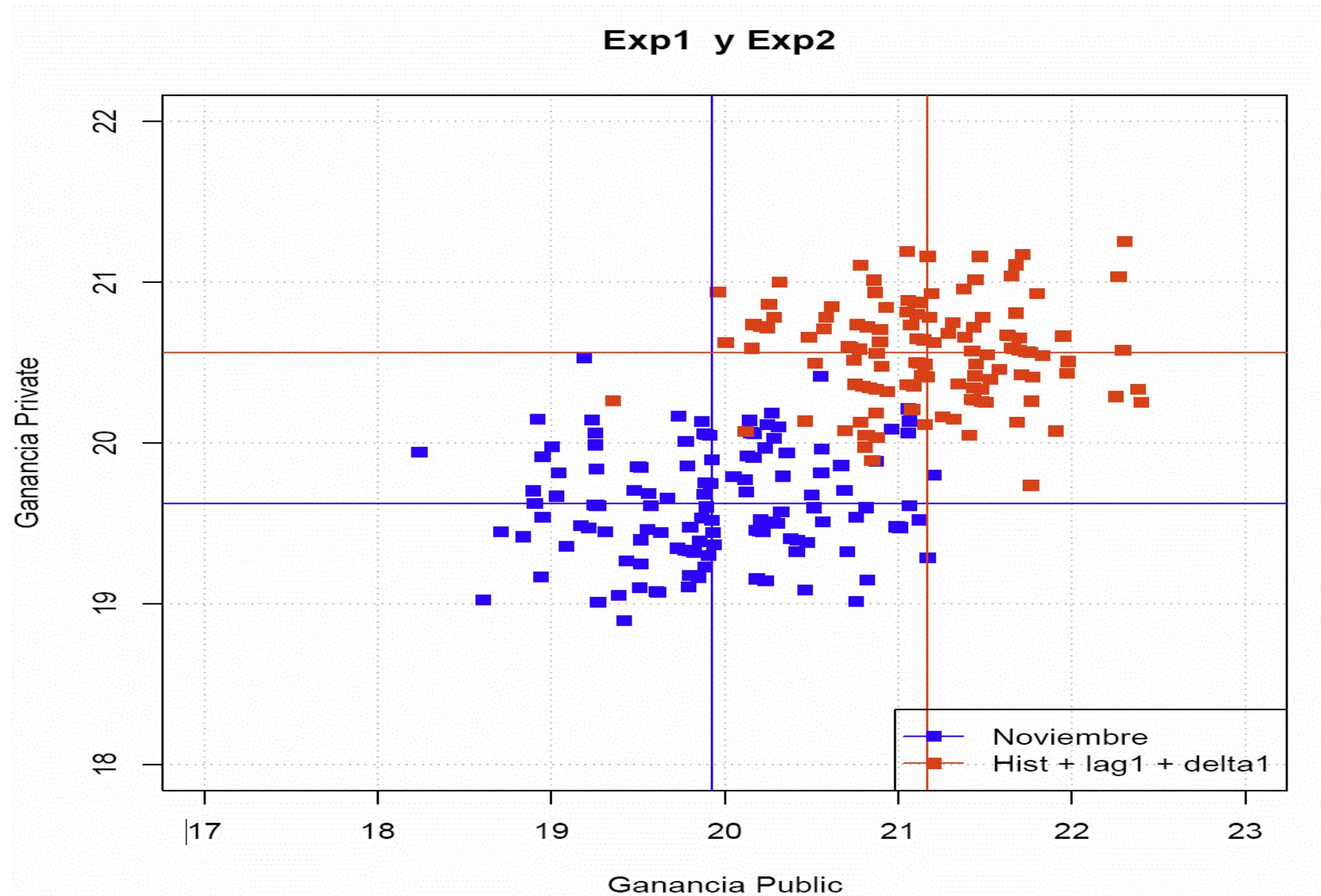
→ 0.032

0.032 es menor que 0.05, SON DISTINTAS

qty	modelo1 campos orig Gan Public	modelo2 lag1 + delta1 GanPublic	Wilcoxon p-value
1	20.55	21.95	1.000
2	19.78	20.29	0.667
3	20.76	20.87	0.400
4	18.95	21.05	0.114
5	19.62	21.68	0.032
6	20.25	20.87	0.009
7	19.79	21.51	0.002
8	19.24	20.79	0.0006
9	20.31	20.91	0.002
10	18.91	21.44	0.0005
11	20.22	20.59	0.0003
12	20.87	21.32	0.0001
13	19.93	21.48	0.00003

← — 5 valores de cada experimento alcanzan para determinar  $gan(exp2) > gan(exp1)$

# Experimento 2 vs 1



Pero este no es el fin de la historia !

Porque aunque el modelo entrenado en datos históricos con el feature engineering de lag1 + delta1 es superior, aún puedo obtener de 19.74 a 21.25 en el Private Leaderboard

Estoy expuesto a demasiada variabilidad

La teoría dice que debo ensamblar modelos lo más distintos posibles, evitar la endogamia

# A Gentle Introduction to Ensemble Diversity for Machine Learning

by [Jason Brownlee](#) on May 14, 2021 in Ensemble Learning

 Tweet

 Tweet

 Share

 Share

Ensemble learning combines the predictions from machine learning models for classification and regression.

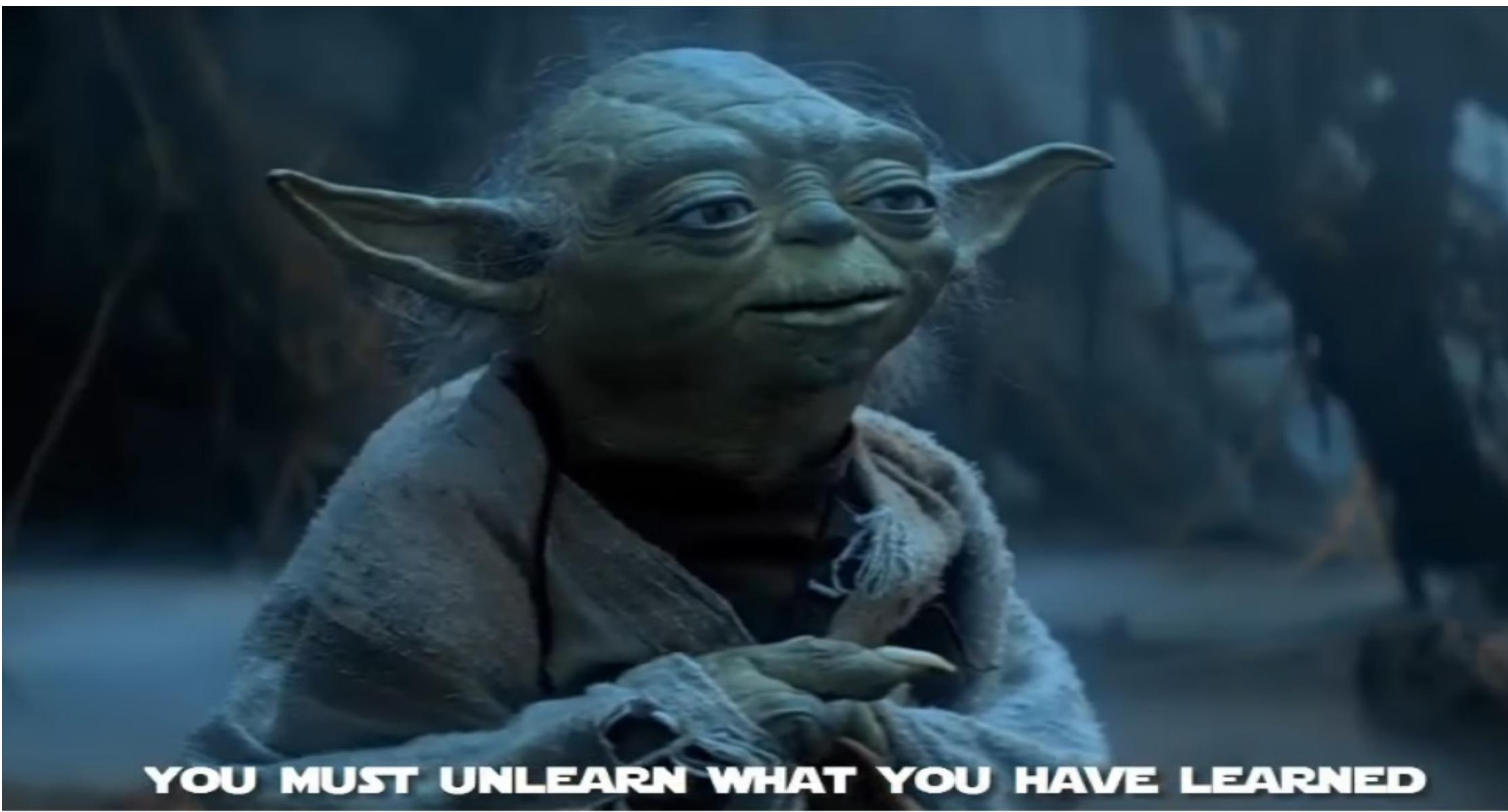
We pursue using ensemble methods to achieve **improved predictive performance**, and it is this improvement over any of the contributing models that defines whether an ensemble is good or not.

A property that is present in a good ensemble is the diversity of the predictions made by contributing models. Diversity is a slippery concept as it has not been precisely defined; nevertheless, it provides a useful practical heuristic for designing good ensemble models.

In this post, you will discover **ensemble diversity** in machine learning.

After reading this post, you will know:

- A good ensemble is one that has better performance than any contributing model.
- Ensemble diversity is a property of a good ensemble where contributing models make different errors for the same input.
- Seeking independent models and uncorrelated predictions provides a guide for thinking about and introducing diversity into ensemble models.



**YOU MUST UNLEARN WHAT YOU HAVE LEARNED**

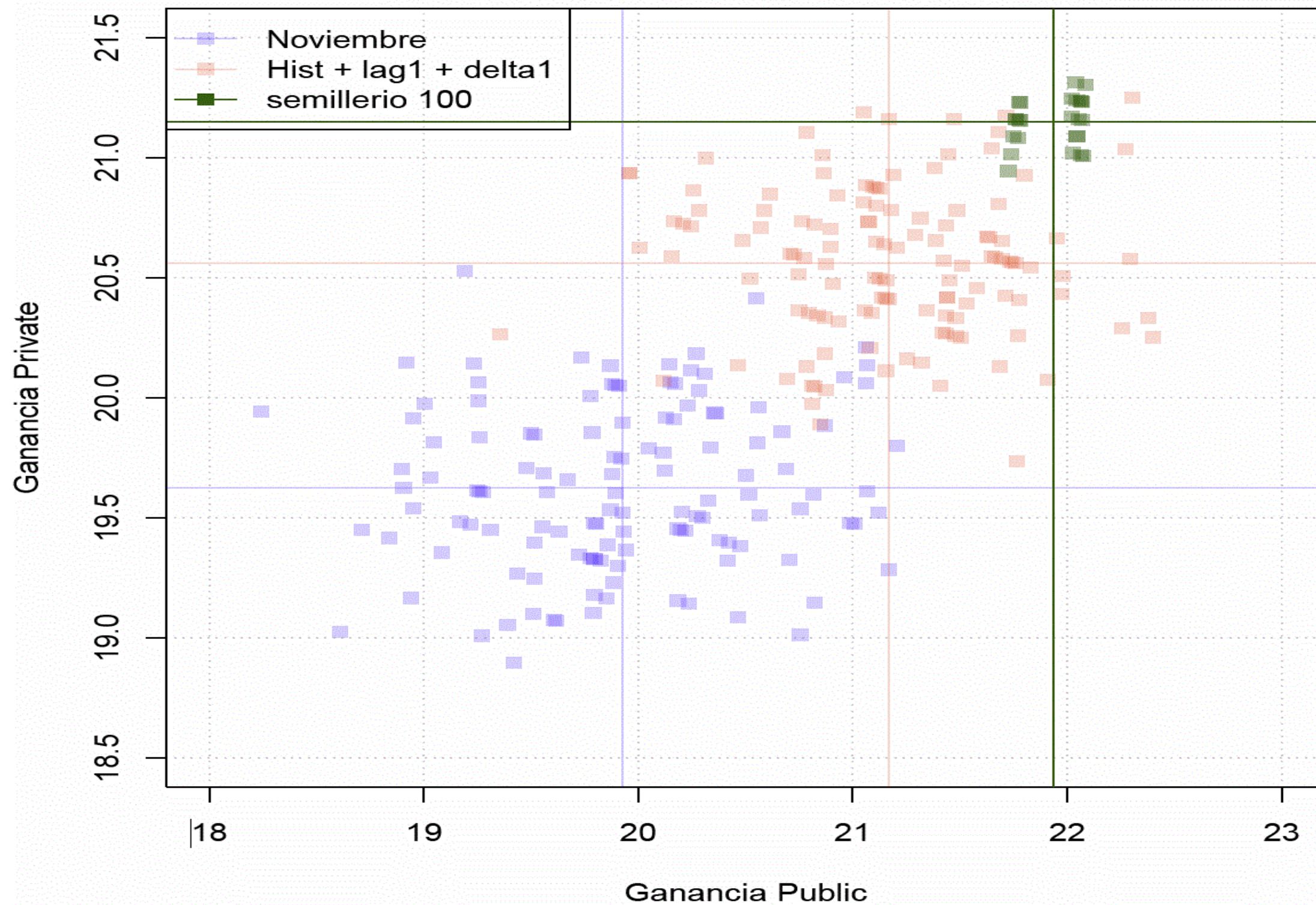
Sin embargo, como la naturaleza de esta asignatura es *cuestionar ideas que otros han aceptado como ciertas*, viene un experimento que no se nos puede negar, en lugar de calcular la ganancia para cada modelo y encomendarme a la "suerte" ya que ni Testing/5-fold cross validation ni el Public Leaderboard son una señal

Ensemble *Semillerios*

Entreno con los mismos hiperparámetros óptimos,  
cambiando solamente la semilla.

Y PROMEDIO las probabilidades que devuelven los  
modelos

# Exp1 , Exp2, Semillerio 100 de Exp2

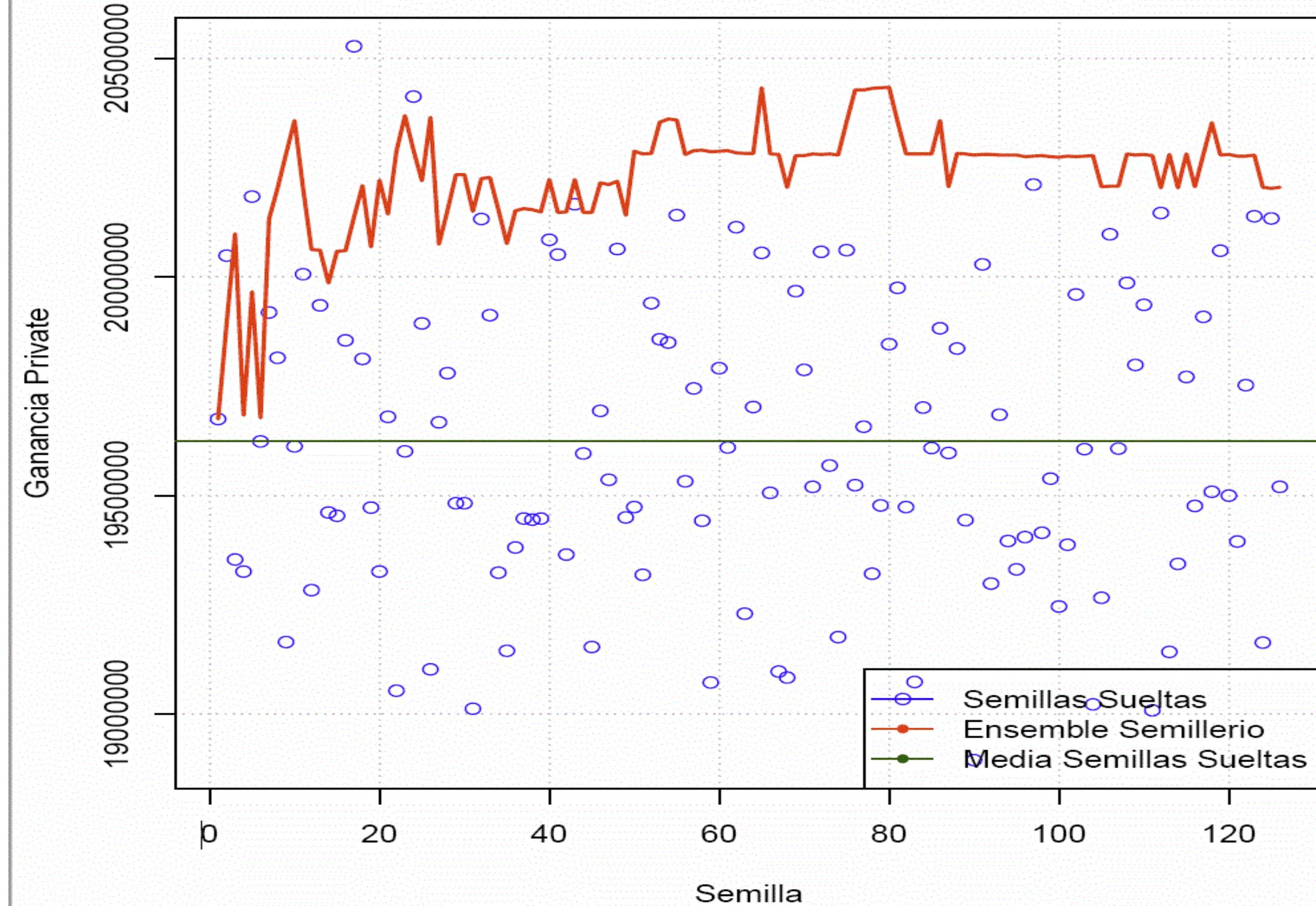


# Efectos de la ensemble de modelo final **promediando** sus probabilidades Private Leaderboard

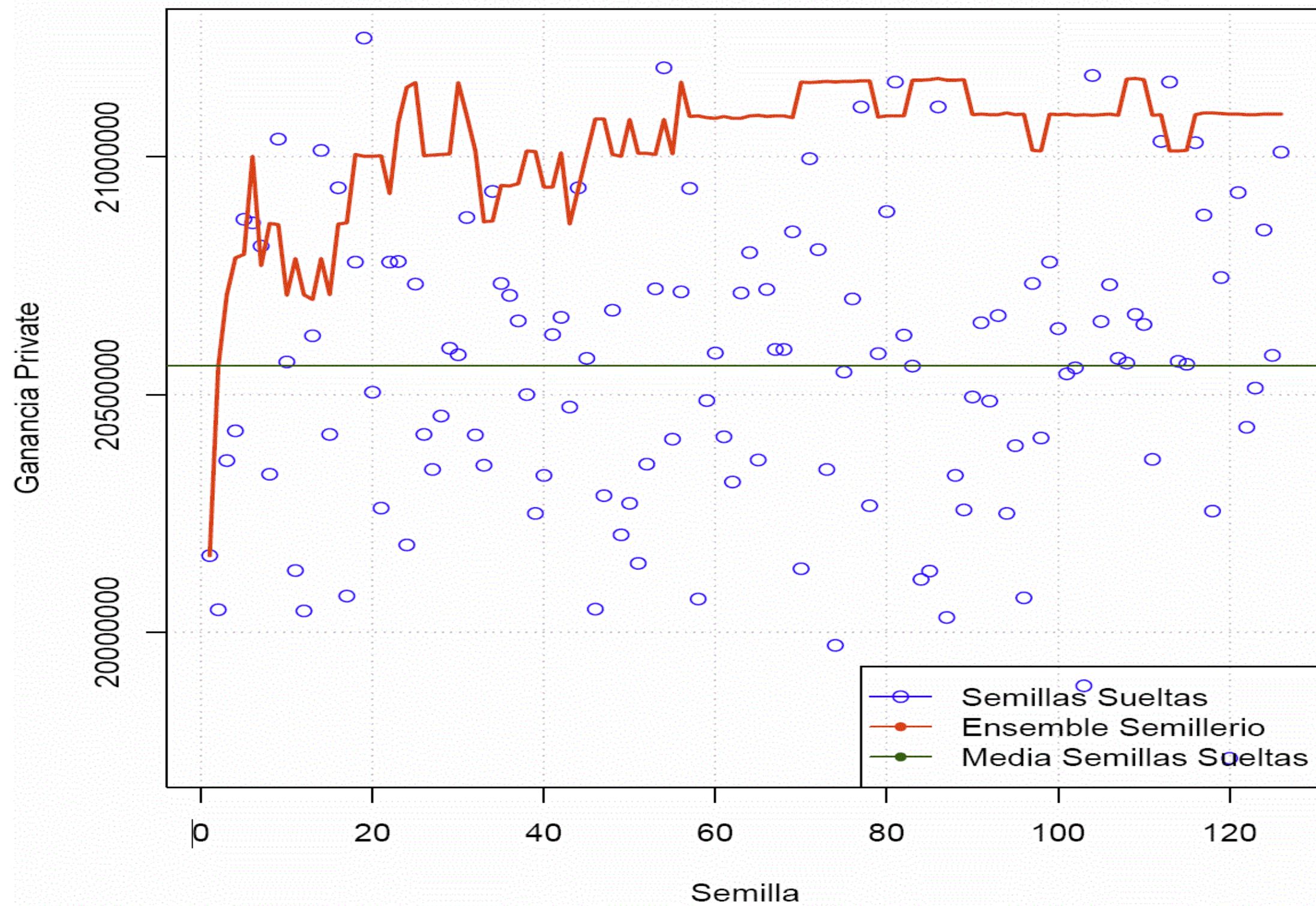
modelos acumulados	dataset noviembre campos originales		dataset 10 meses lag1 + delta1	
	ganancia media	desvio estandar	ganancia media	desvio estandar
1	19.64	0.35	20.7	0.27
5	20.04	0.23	21.0	0.19
10	20.09	0.20	21.1	0.15
20	20.19	0.16	21.1	0.10
50	20.21	0.10	21.2	0.06

\$ 500k adicionales en Experimento DOS , solo con semillerio, nada mal ...

# Experimento 1 Semillerio 11000 envios



## Experimento DOS Semillerio 11000 envios



Los puntos verdes,  
*Semillerio-100 de Modelo Power*  
no solo tiene mayor ganacia promedio  
que *Modelo Power*  
sino que además, sensiblemente menor varianza

Ya no van a haber sorpresas en datos nuevos  
No van a haber caídas ni subidas tan pronunciadas en el Private  
Leaderboard

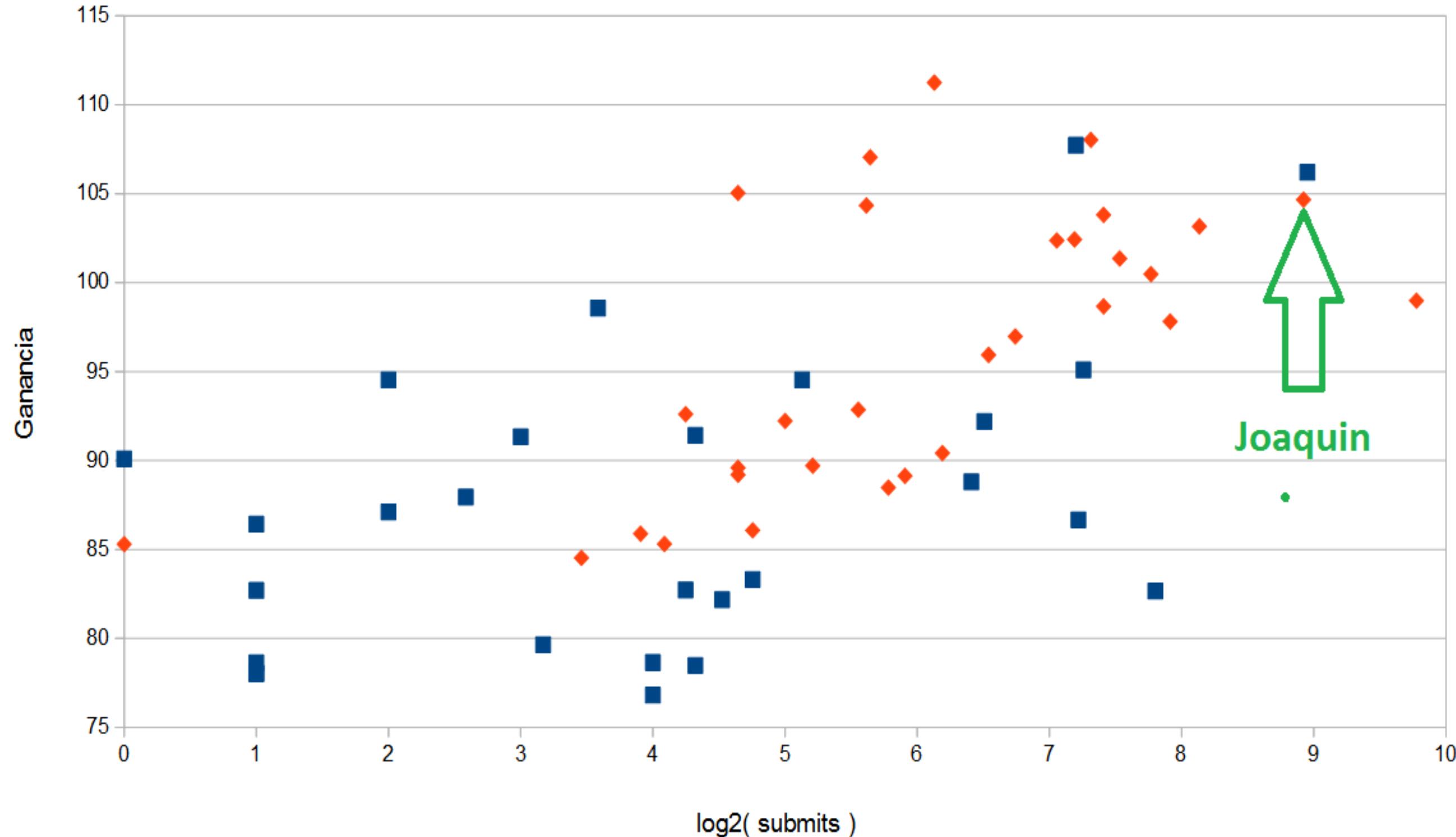
Si  $X_1, X_2, \dots, X_n, \dots$  son muestras aleatorias extraídas de una población con media global  $\mu$  y varianza finita  $\sigma^2$ , y si  $\bar{X}_n$  es la media muestral de las primeras  $n$  muestras, entonces la forma límite de la distribución,  $Z = \lim_{n \rightarrow \infty} \left( \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}}} \right)$ , con  $\sigma_{\bar{X}} = \sigma / \sqrt{n}$ , es una distribución normal estándar.<sup>5</sup>

No era nada intuitivo  
que generar un ensemble de LightGBM's  
cambiando solo la semilla  
iba a generar un modelo superador

Incluso, es una idea tan hereje que nadie en su sano juicio  
perdía tiempo en probarla !

¿Cómo se ve el semillerio  
en la Primera Competencia Kaggle?

## Competencia UNO Ganancia vs Submits



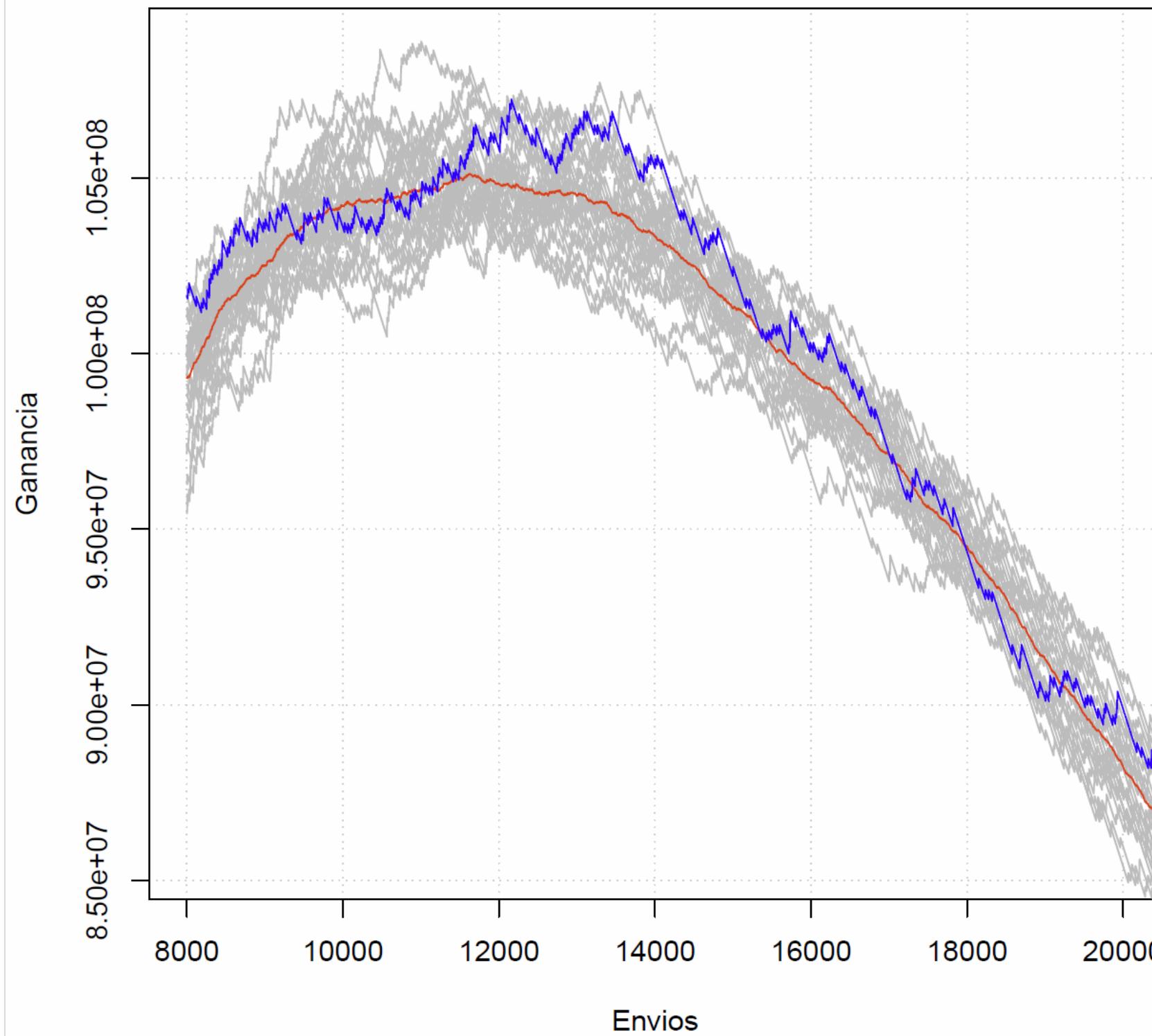
# La solución de Joaquín Tschopp

mejor iteracion Bayesian Optimization	
num_iterations	2215
learning_rate	0.021 Razonable !
num_leaves	22
min_data_in_leaf	912
feature_fraction	0.34
extra_trees	FALSE

# Hago un **30-semillero** (30 semillas)

de la mejor iteración  
de Joaquin Tschopp

repeticion= 1, mejor gan prom = 106216261



Un n-semillerio es un animal de una nueva especie

Es un Random Forest de LightGBM's

# Los hiperparámetros óptimos de un LightGBM “solito”

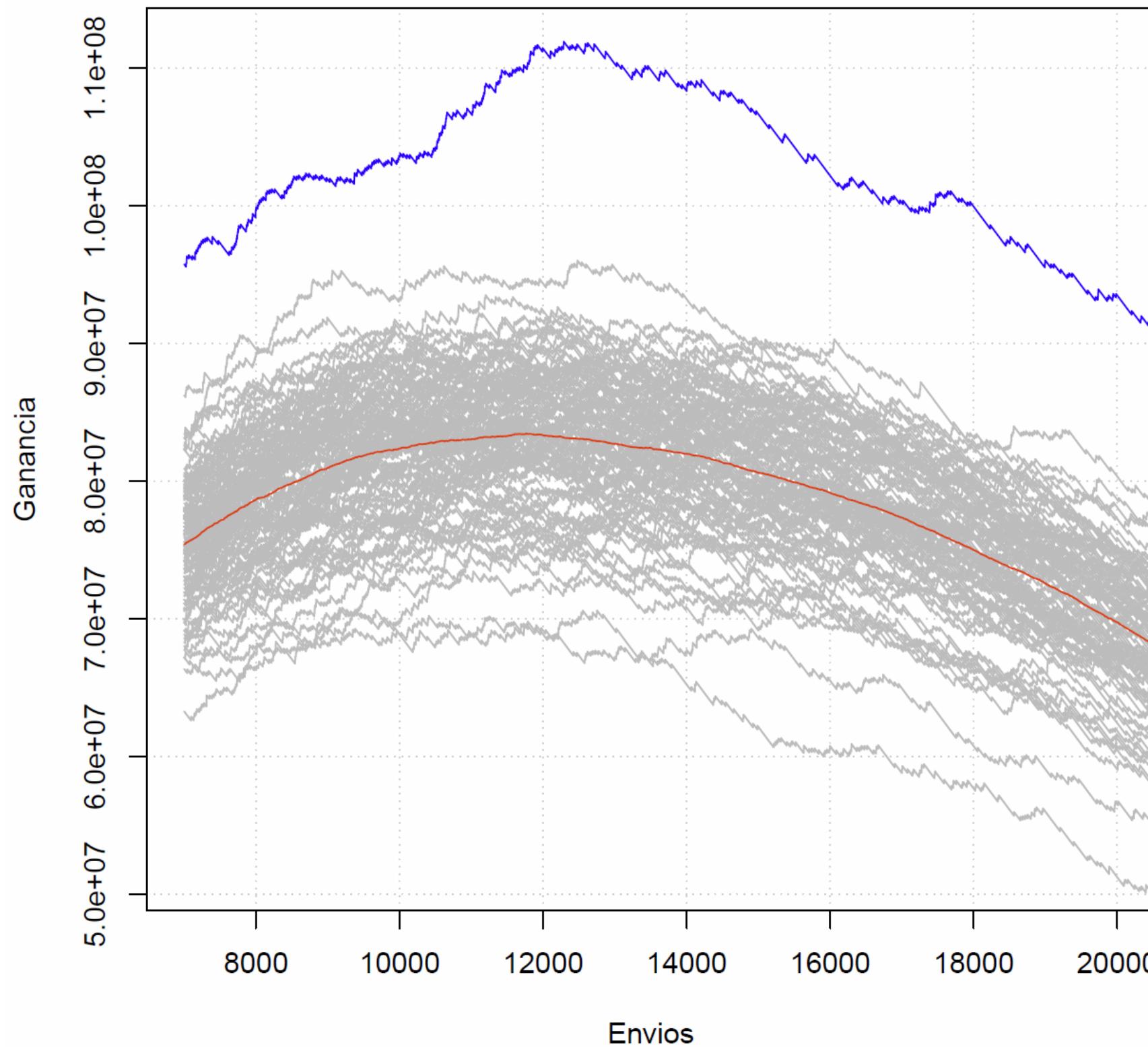
¿son los hiperparámetros óptimos que debe tener  
un LightGBM cuando forma parte de un semillero?

## Parámetros inesperados 100-semillerio

num_iterations	23	
learning_rate	<b>0.34</b>	Inesperado
num_leaves	32	
min_data_in_leaf	711	
feature_fraction	0.20	
extra_trees	FALSE	

Cada LightGBM en forma individual es muy débil, pero formando parte de un semillerio ...

repeticion= 1, mejor gan prom = 110719171



Public = 107.92 M  
Private = 113.15 M

Joaquin Tschopp  
si hubieras usado semillerio  
quedabas **1ro**  
en la Primera Competencia

a pesar que Guillermo Piazza  
partió de un mejor dataset

( principalmente hizo un correcto <train,validate,test> )

	Public	Private	Total
Joaquin	104.367	104.671	104.580
Guillermo	94.614	111.251	106.260
Semillerio (dataset Joaquín)	107.916	113.150	111.580

Joaquín	Guillermo
<p>Train= {202102, 202103} 0.75 undersampling</p> <p>Validate= {202104}</p> <p>Test= {202104}</p>	<p>Train= {202102} sin undersampling</p> <p>Validate= {202103}</p> <p>Test= {202104}</p>
final_train= {202102, 202103, 202104}	final_train= {202102, 202103, 202104}
<p>Con lag1 y delta1</p> <p>Con variables manuales</p> <p>443 atributos</p>	<p>Con lag1 y delta1</p> <p>Con variables manuales</p> <p>581 atributos</p>

# Disgresión técnica para Joaquin

Nunca, pero más en tu vida  
elimines una variable de esta forma

```
dataset <- dataset[, -c("mpayroll")]
```

Hacelo así

```
dataset[, mpayroll := NULL ]
```

The End