

Clase de Cierre

ciclo 2021

Laboratorio de Implementación I

Maestría en Ciencia de Datos

Universidad Austral

Agenda 26-oct-2021 "dos videos"

- Motivación
- Enfoque Estadístico del Aprendizaje Automático
- Planteo del problema
- Resultados Experimentales
- Tests Estadísticos para la comparación de Modelos Predictivos



- Enfoque pragmático filosofía Machine Learning a Laboratorio de Implementación I

Motivación

Motivación

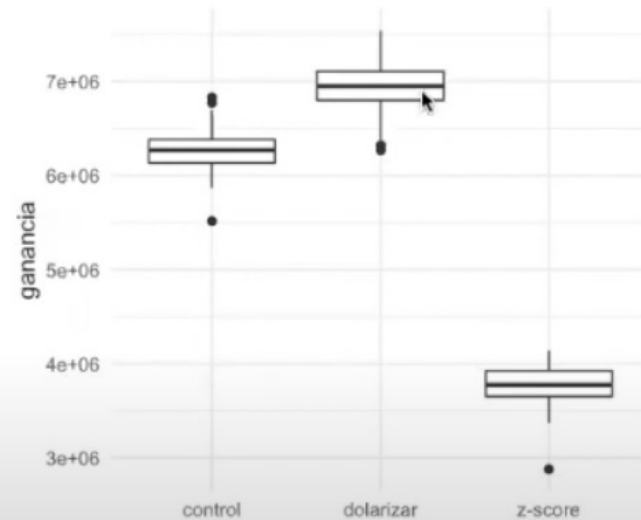
Observación en la última clase presencial

Un 15% de los alumnos refiere que están utilizando LightGBM porque les funciona mejor que Stacking, mientras que el 82% refiere que Stacking funciona mejor.

John Sandoval refiere que su votación manual de tres modelos en Excel supera a TODO lo que ha seguido probando, incluyendo a Stacking.

resultados

normalizar vs dolarizar



Normalizar tiene efectos catastróficos sobre la predicción. Dolarizar (en las condiciones testeadas), genera una modesta mejora en la ganancia.

Germán Beldorati Stark, 29 años, biólogo,
gerbeldo 4° puesto Private Leaderboard Kaggle
trabajó **dolarizando** los atributos monetarios del dataset.



En primer lugar, hice un ajuste por inflación de las variables nominales en pesos, utilizando el IPC del INDEC. Si bien esto aumentó la ganancia en el log (1 fold de Nov-2020 no lo hizo es el Leaderboard Público de Kaggle. Sino que por el contrario la ganancia en el Public fue mucho menor. Por esta razón abandoné este camino.

Daiana Belén Sparta, 33 años, economista

Daiana Sparta 6° puesto Private Leaderboard Kaggle

NO deflactó los atributos monetarios del dataset porque

"la ganancia en el Public fue mucho menor"

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**." Pablo Tomás Barco

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

Motivación comentarios

María del Pilar Arias propone

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

Sebastián Spena , del equipo ganador, responde

"Si supiera, te lo diría."

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

Nicolás Pereyra Zorraquin , del equipo ganador

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"

Lucas Ibañez

Motivación comentarios

"... no puedo conectar todo el análisis hecho con los resultados de kaggle." María Belén Bernatene

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

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

Motivación

Nombres de equipos

- Monos que apretan palancas
- Team Suerte y Overfitting



Resultados Experimentales

Experimento 1

¿Cuánto puede variar la ganancia de un modelo?

¿Qué relación hay entre la ganancias medidas en los tres datasets Testing, Public y Private?

¿ Sin un modelo M1 da más ganancia que M2 en testing, también es mejor en el Public ? ¿ Y en el Private?

Experimentos

En el repositorio GitHub de la materia <https://github.com/labo-imp/labo2021> está la nueva carpeta **ClaseCierre** donde se encuentran los scripts y los resultados de los experimentos

- **R** carpeta *scripts*
- **work** carpeta *resultados consolidados*

todos los scripts deben correr en la nube, partiendo del dataset original [paquete_premium.csv.gz](#)

Experimento 1 Objetivo

Objetivo: analizar la variabilidad de un modelo *fijo*

A partir del dataset original en donde solo se corrigen variables *rotas*, se buscan los hiperparámetros

óptimos del LightGBM con una Optimización

Bayesiana, train=[201901,202010] test=[202011]

Finalmente, se observa el comportamiento de

regenerar el modelo con distintas semillas en:

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

Experimento 1 dataset

Utilizando el script `951_dataset_epic.r` se genera el dataset `dataset_epic_v951.csv.gz`

La única palanca que se activa es

```
palancas$corregir <- TRUE
```

que llama a la función `Corregir(dataset)`

básicamente lo que realiza es asignar `NA` a las variables que para ese mes el sector de DataWarehousing cometió un gravísimo error y asignó casi todos los valores en cero.

Experimento 1 dataset

En Experimento 1 intencionalmente **NO** se crea ninguna variable nueva, ni en el mismo mes ni tampoco histórica.

Experimento 1 Optimización Bayesiana

Se realiza una Optimización Bayesiana utilizando el script `961_epic.r` (copia del script `822_epic.r`) en donde

training: `[201901, 202010]` 22 meses

validation: la primera mitad de `[202011]`

testing: la segunda mitad de `[202011]`

subsampling: 10% de los "CONTINUA"

Experimento 1 Optimización Bayesiana

LightGBM hiperparámetros óptimos	
learning_rate	0.0689581204
feature_fraction	0.4820239822
min_data_in_leaf	1379
num_leaves	119
num_iterations	173
ratio_corte	0.0461216724

Resultados	
Testing	7,272,500
Public	24.20477
Private	21.83548

Experimento 1 jugando con la semilla

LightGBM no es un algoritmo siempre determinístico (por ejemplo cuando `feature_fraction < 1`), para lo cual utiliza una semilla, que hasta ahora siempre hemos dejado fija

```
param_basicos <- list( objective= "binary",  
                        metric= "custom",  
                        first_metric_only= TRUE,  
                        boost_from_average= TRUE,  
                        feature_pre_filter= FALSE,  
                        verbosity= -100,  
                        seed= 999983
```

Experimento 1 jugando con la semilla

Pregunta de Investigación:

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

¿Cuál es la variabilidad inherente de un modelo generado con LightGBM ?

Experimento 1 jugando con la semilla

Scripts [981_semillerio.r](#) y [991_semillerio_kaggle.r](#)

```
#me genero un vector de semilla buscando primos
```

```
primos <- generate_primes(min=100000, max=1000000)
```

```
#genero TODOS los numeros primos entre 100k y 1M
```

```
ksemillas <- sample(primos)[ 1:CANTIDAD_SEMILLAS ]
```

```
#me quedo con CANTIDAD_SEMILLAS primos al azar
```

```
ksemillas <- c( 999983, ksemillas )
```

```
for( semillita in ksemillas ) #itero por las semillas  
{
```

```
  gc()
```

```
  param_completo$seed <- semillita #asigno la semilla a esta corrida
```

Experimento 1 jugando con la semilla

En el script `981_semillerio.r` se entrena en [201901, 202009] y se mide la ganancia en todo [202011]

En el script `991_semillerio_kaggle.r` se entrena en [201901, 202011] y se mide la ganancia en Kaggle

No se utiliza undersampling en ninguno de los dos scripts. Siempre se elimina el dañado mes [202006]

Para este caso (no hay feature engineering) se generan 500 modelos, utilizando 500 semillas.

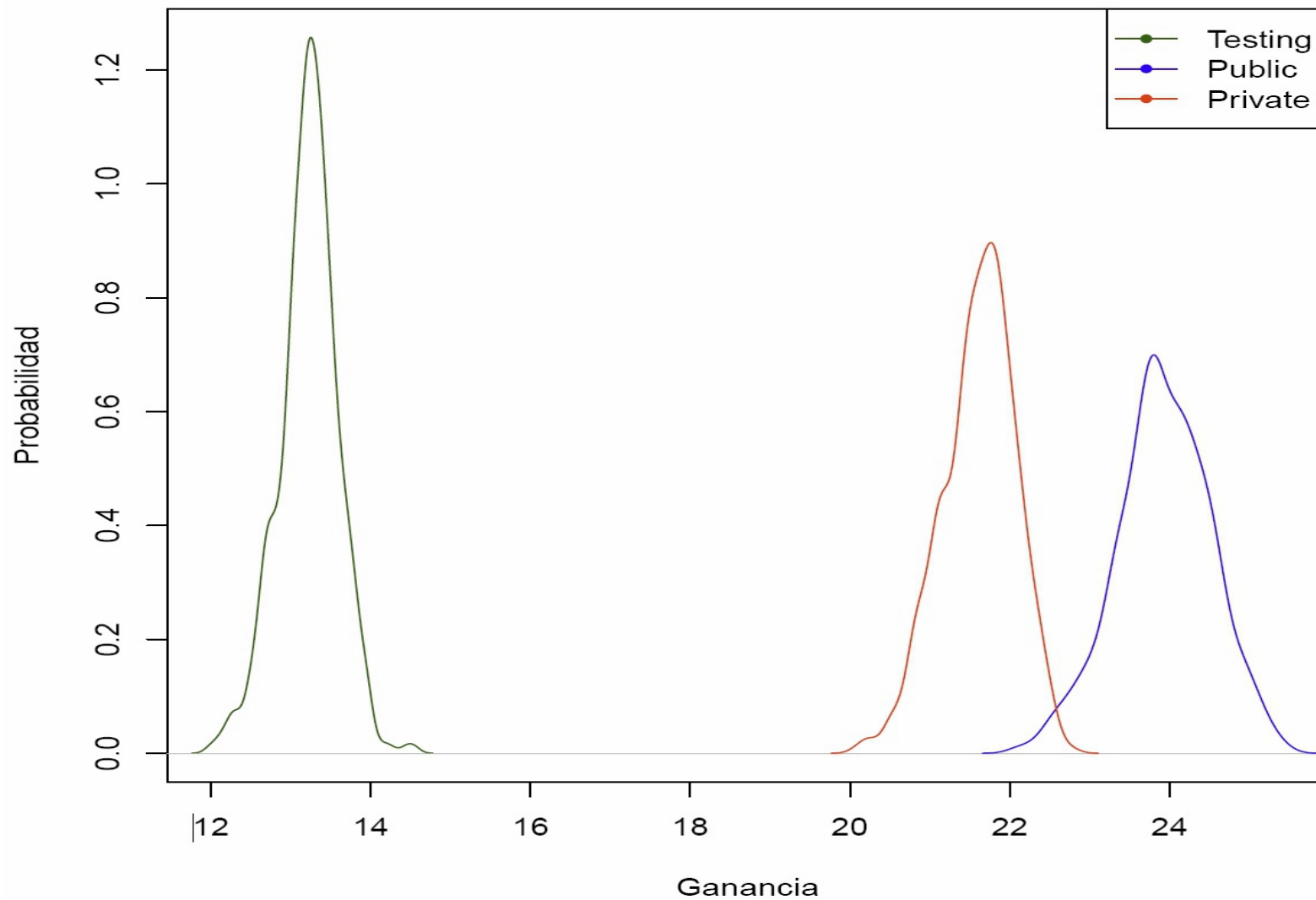
Experimento 1 resultados

Los resultados del Experimento 1 que se se obtienen con el script [921_experimento.r](#)

Cambiando las semillas las corridas jamás dan la misma ganancia ni en el Testing, 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

Densidades 500 puntos

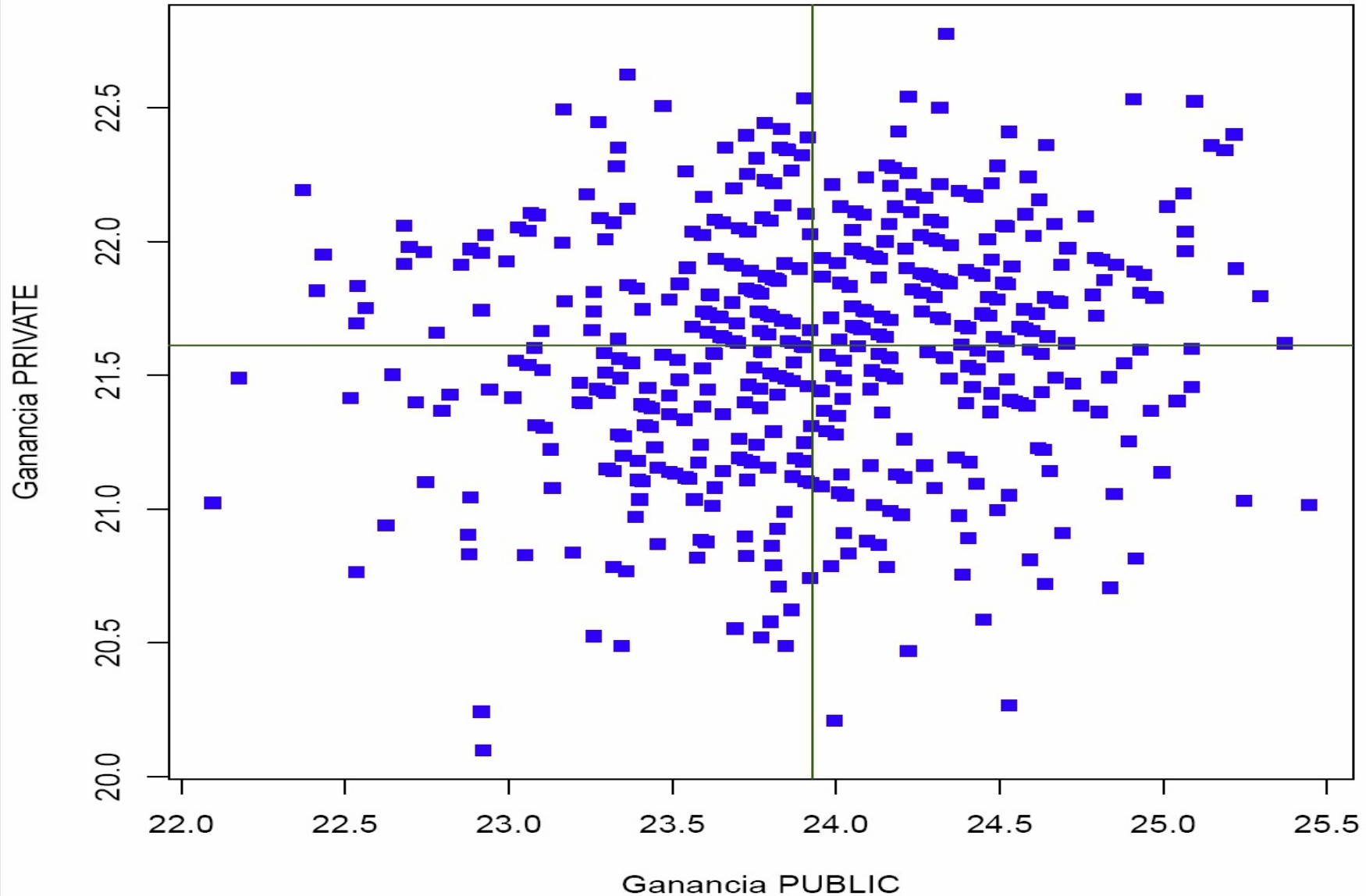


Experimento 1 resultados

Ganancia	mean	sd
Testing	13.2	0.368
Public	23.9	0.590
Private	21.6	0.459

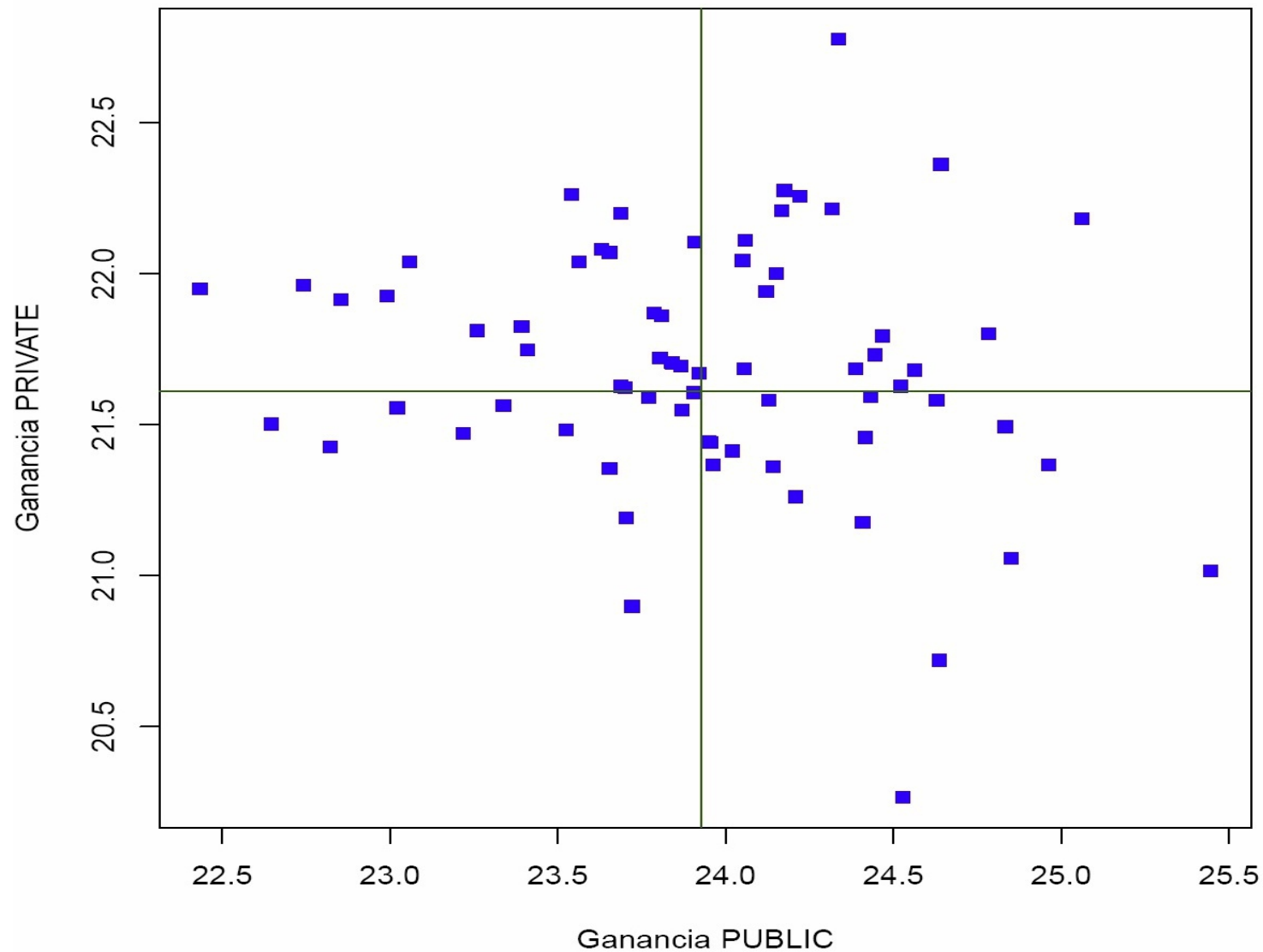
Experimento 1 aleatoriedad

Ganancias Private vs Public puntos 501



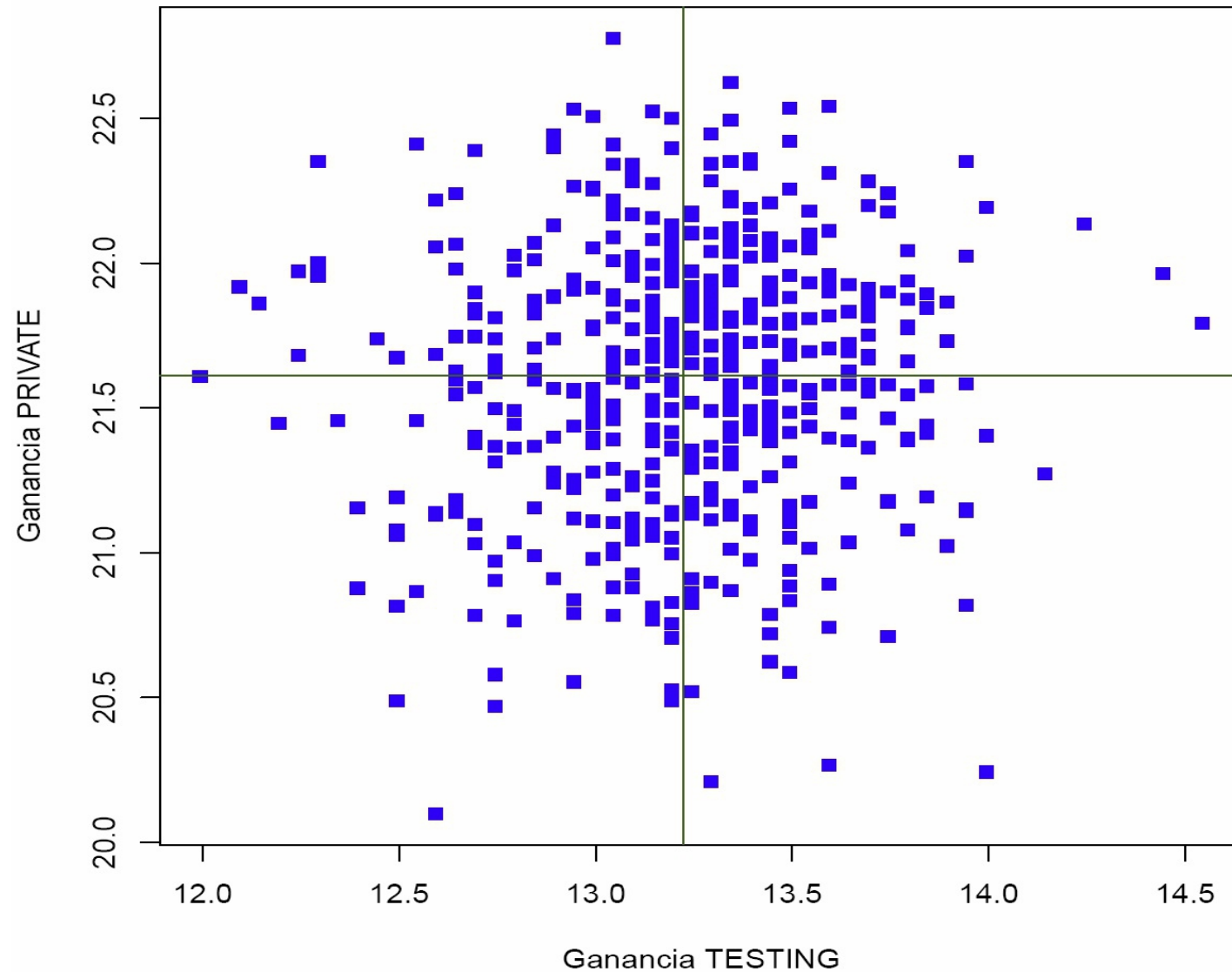
Experimento 1 aleatoriedad

Ganancias Private vs Public 70 puntos



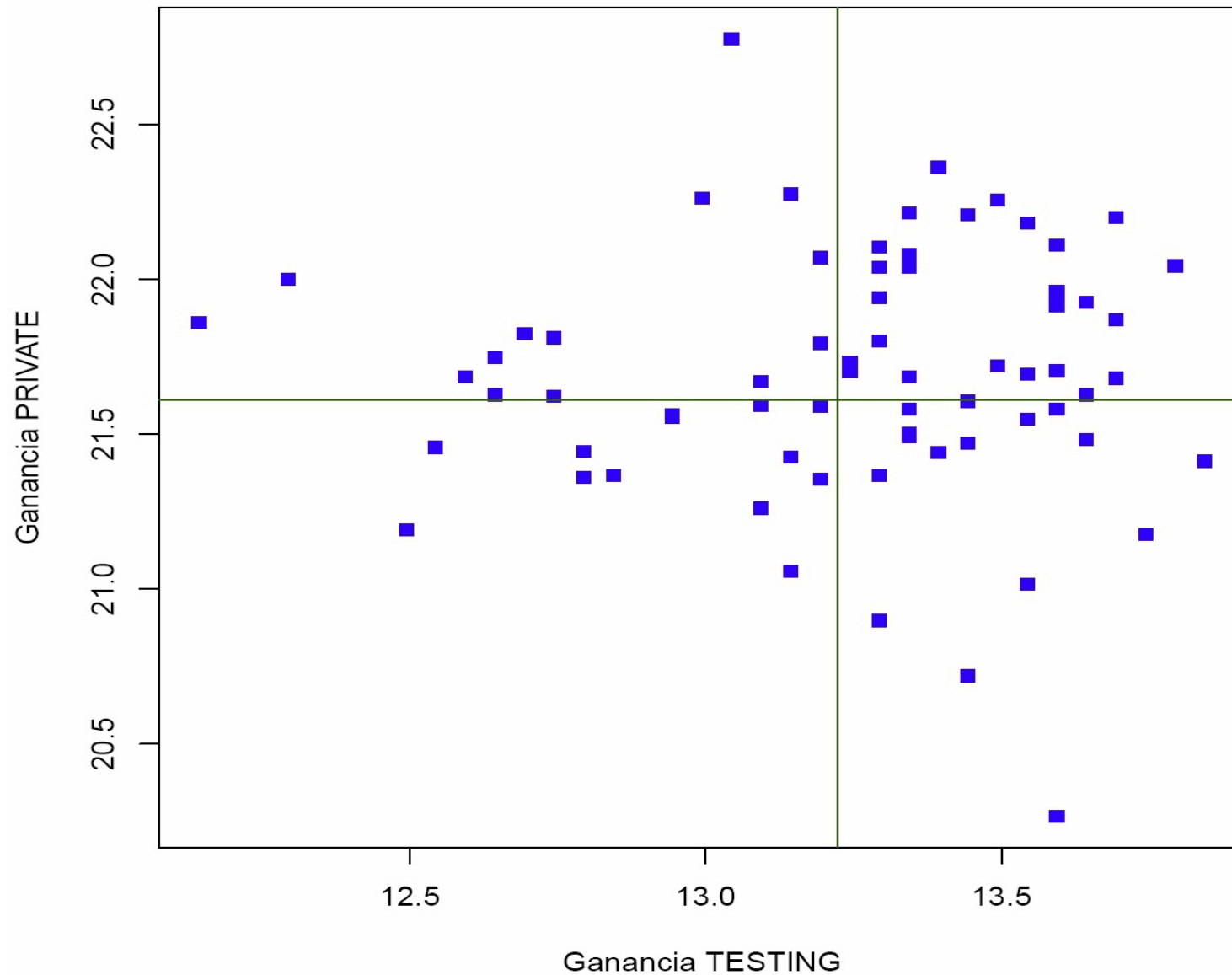
Experimento 1 aleatoriedad

Ganancias Private vs Testing puntos 501



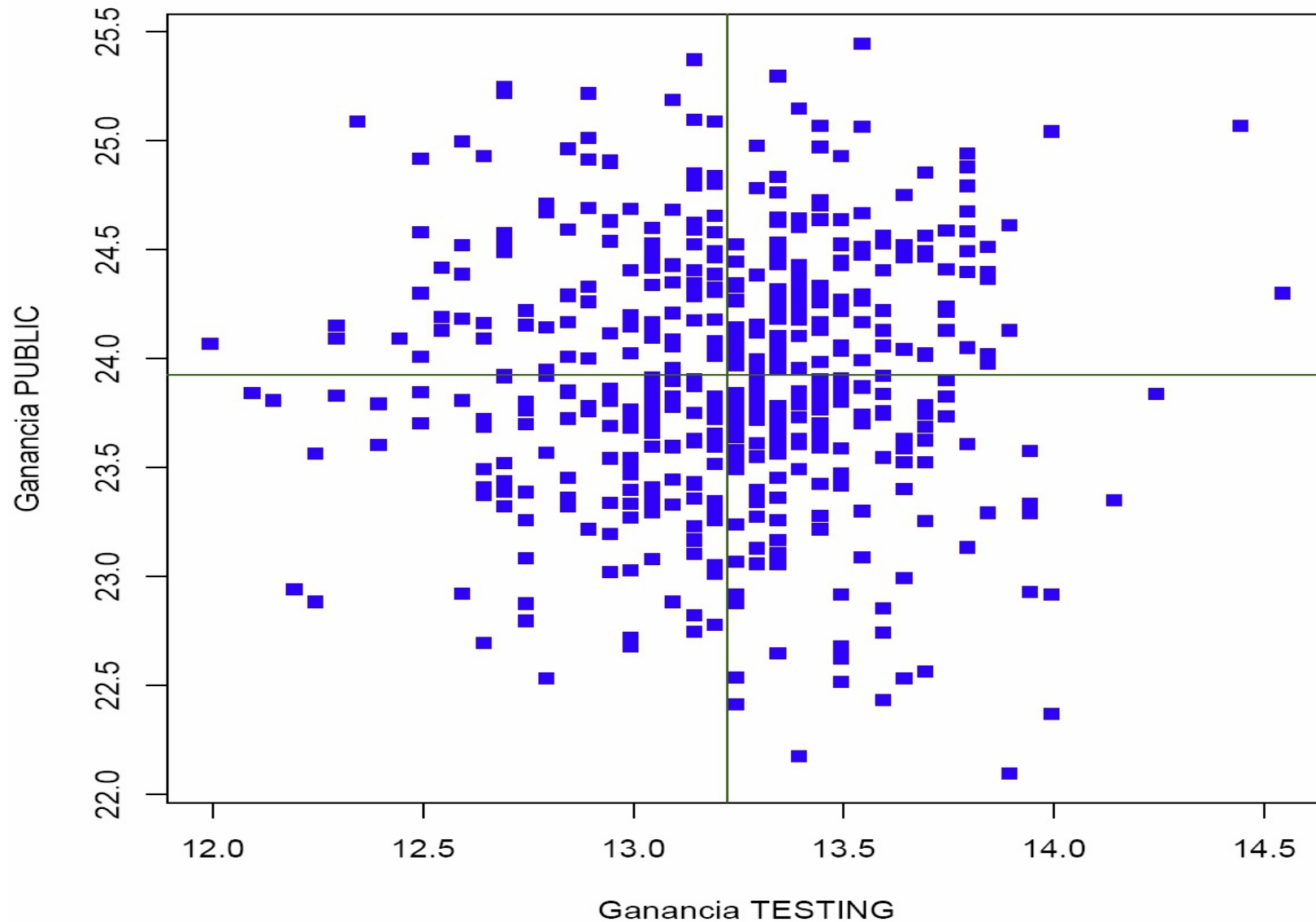
Experimento 1 aleatoriedad

Ganancias Private vs Testing 70 puntos

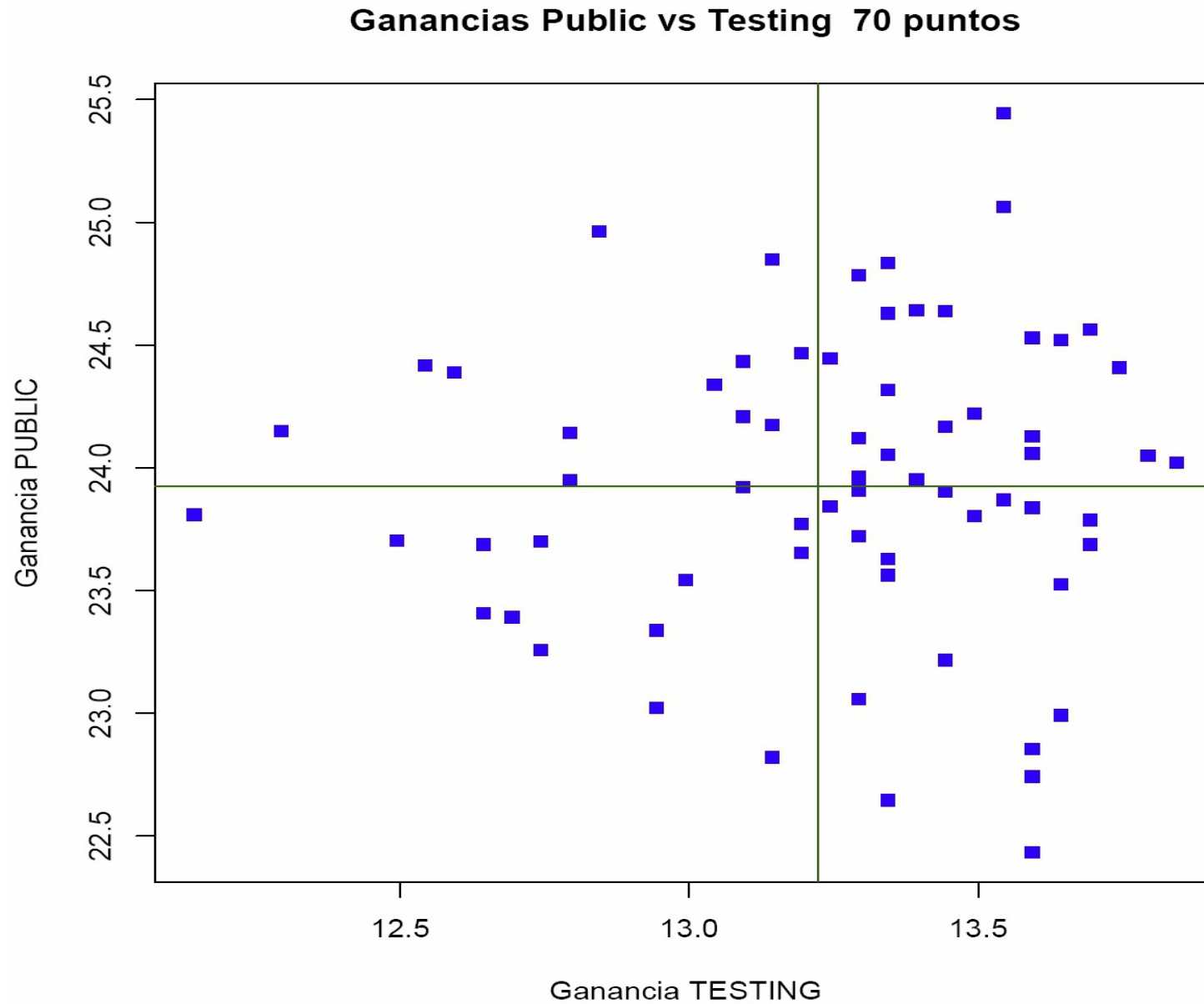


Experimento 1 aleatoriedad

Ganancias Public vs Testing puntos 501



Experimento 1 aleatoriedad



Experimento 1 conclusión

- Los resultados en los datasets de testing, Public y Private poseen una distribución 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, por más que en el otro dataset lo esté.

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

Experimento 2

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=[201901,202010] test=[202011] Finalmente, se observa el comportamiento de **regenerar** el modelo con distintas semillas en:

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

Experimento 2 dataset

Utilizando el script `952_dataset_epic.r` se genera el dataset `dataset_epic_v952.csv.gz`

se activan tres palancas

```
palancas$corregir <- TRUE  
palancas$lag1      <- TRUE  
palancas$delta1    <- TRUE
```

Experimento 2 Optimización Bayesiana

Se realiza una Optimización Bayesiana utilizando el script `962_epic.r` (copia del script `822_epic.r`) en donde

training: `[201901, 202010]` 22 meses

validation: la primera mitad de `[202011]`

testing: la segunda mitad de `[202011]`

subsampling: 10% de los "CONTINUA"

Experimento 2 Optimización Bayesiana

LightGBM hiperparámetros óptimos	
learning_rate	0.0289933062436
feature_fraction	0.9141429986475
min_data_in_leaf	367
num_leaves	455
num_iterations	461
ratio_corte	0.0465659156440

Resultados	
Testing	7,706,250
Public	25.21729
Private	22.56761

Experimento 2 vs 1

Métrica	Variables Originales	Lag 1 + Delta1
Testing	7,272,500	7,706,250
Public	24.20577	25.21729
Private	21.83548	22.56761

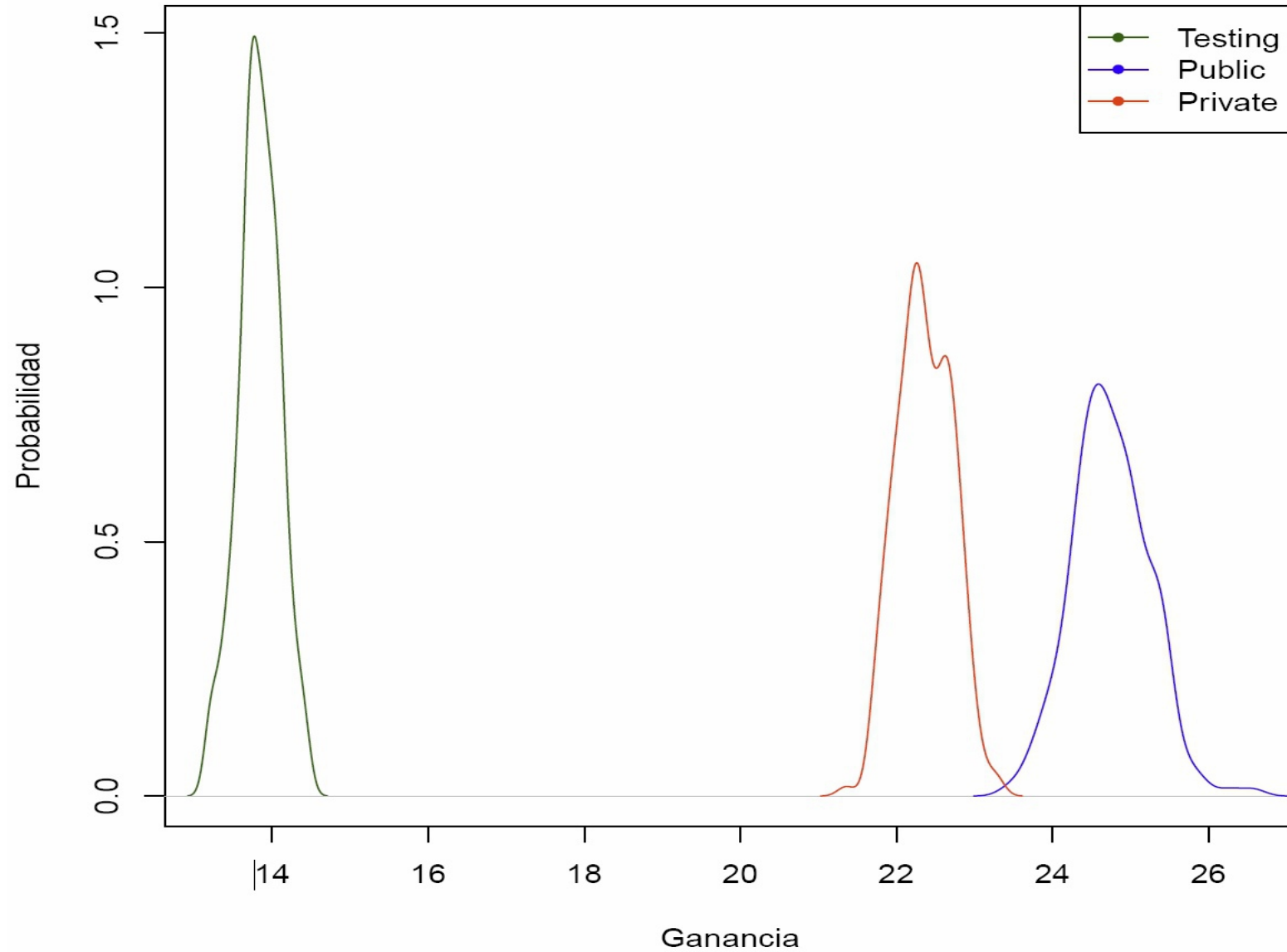
Experimento 2 resultados

Los resultados del Experimento 2 que se obtienen con el script `921_experimento.r`

La comparación con los resultados del Experimento 1 se hacen por medio del script `922_experimentos_compara.r`

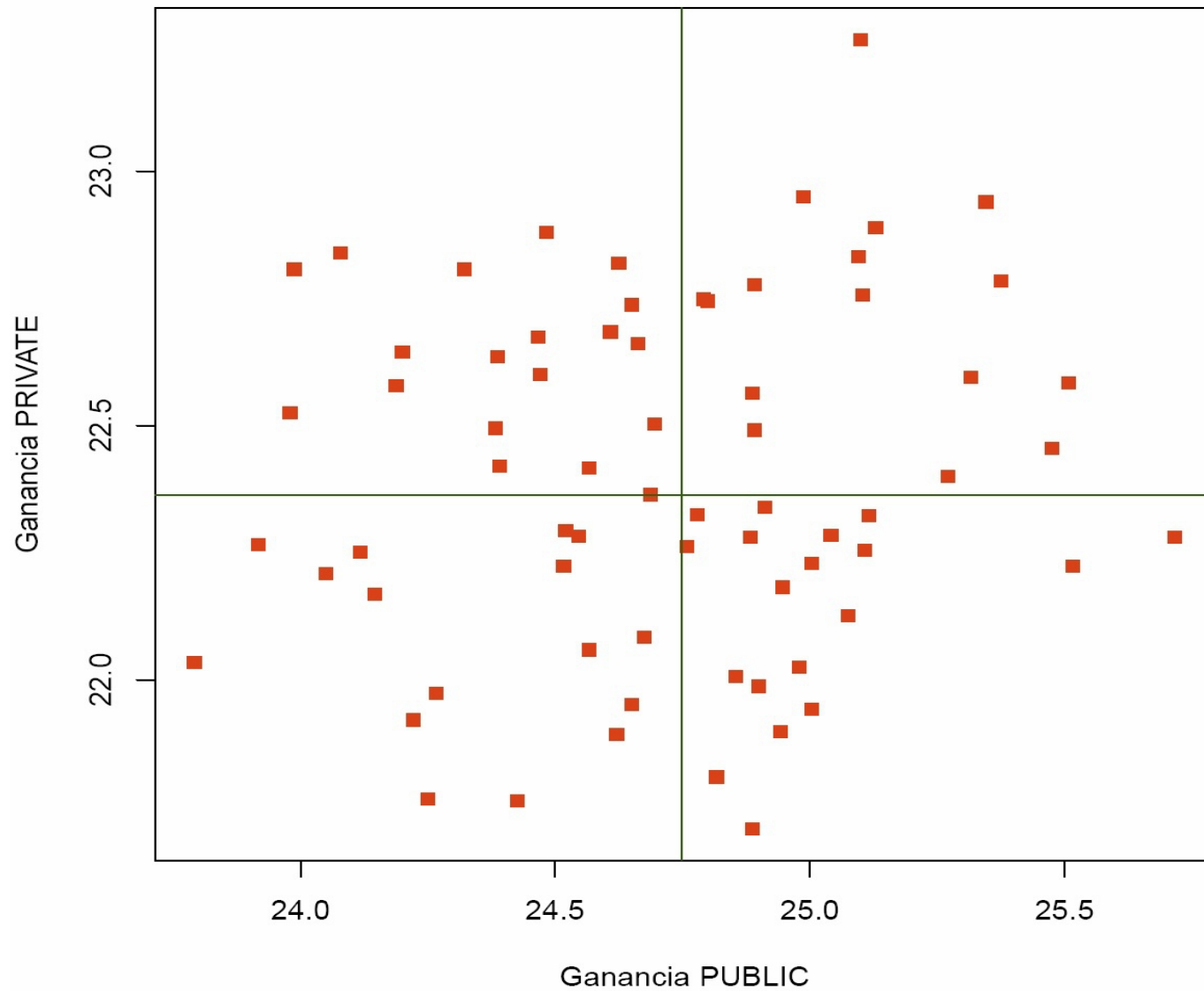
Experimento 2 resultados

Densidades 500 puntos



Experimento 2 aleatoriedad

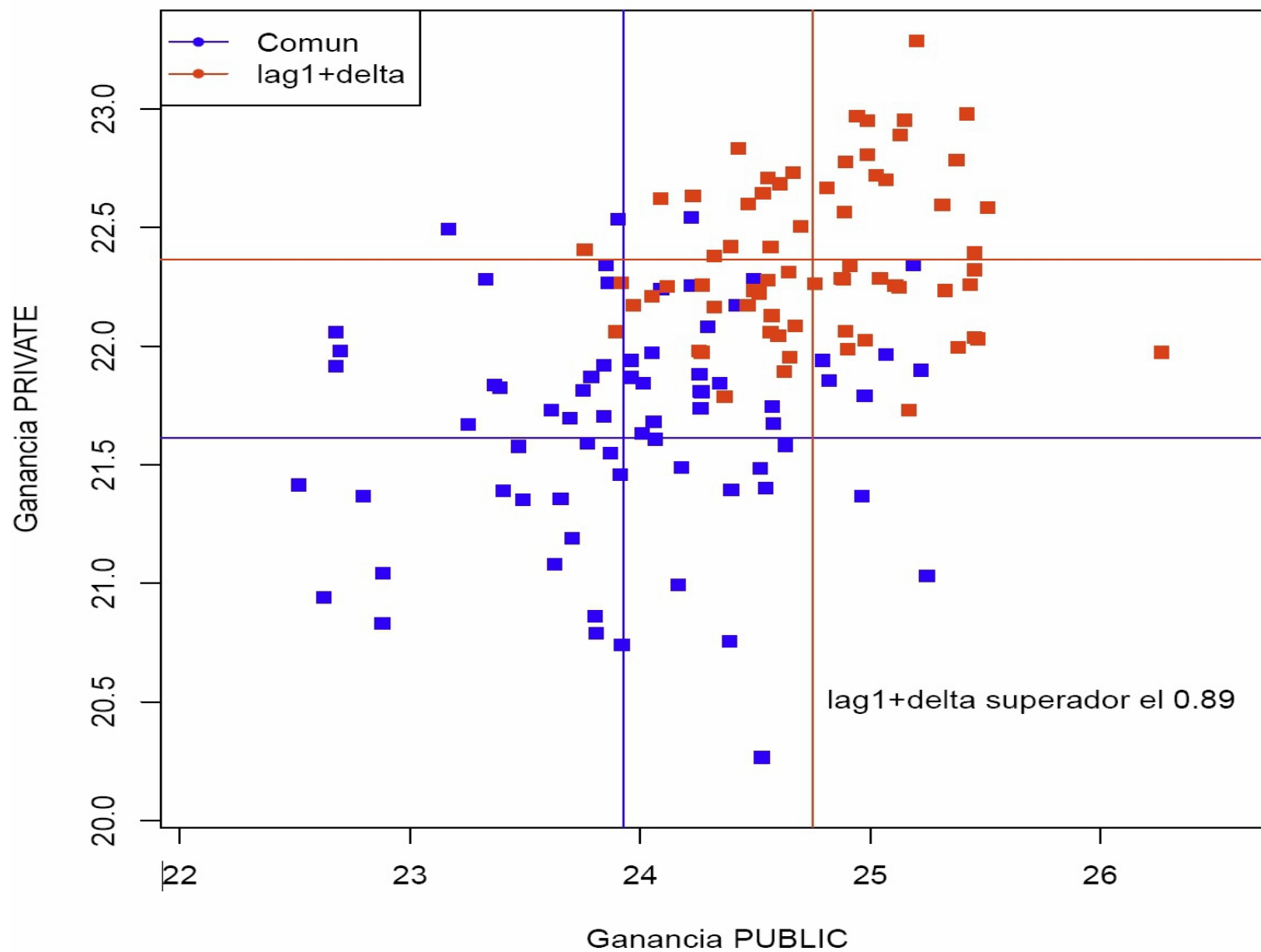
Ganancias Private vs Public 70 puntos



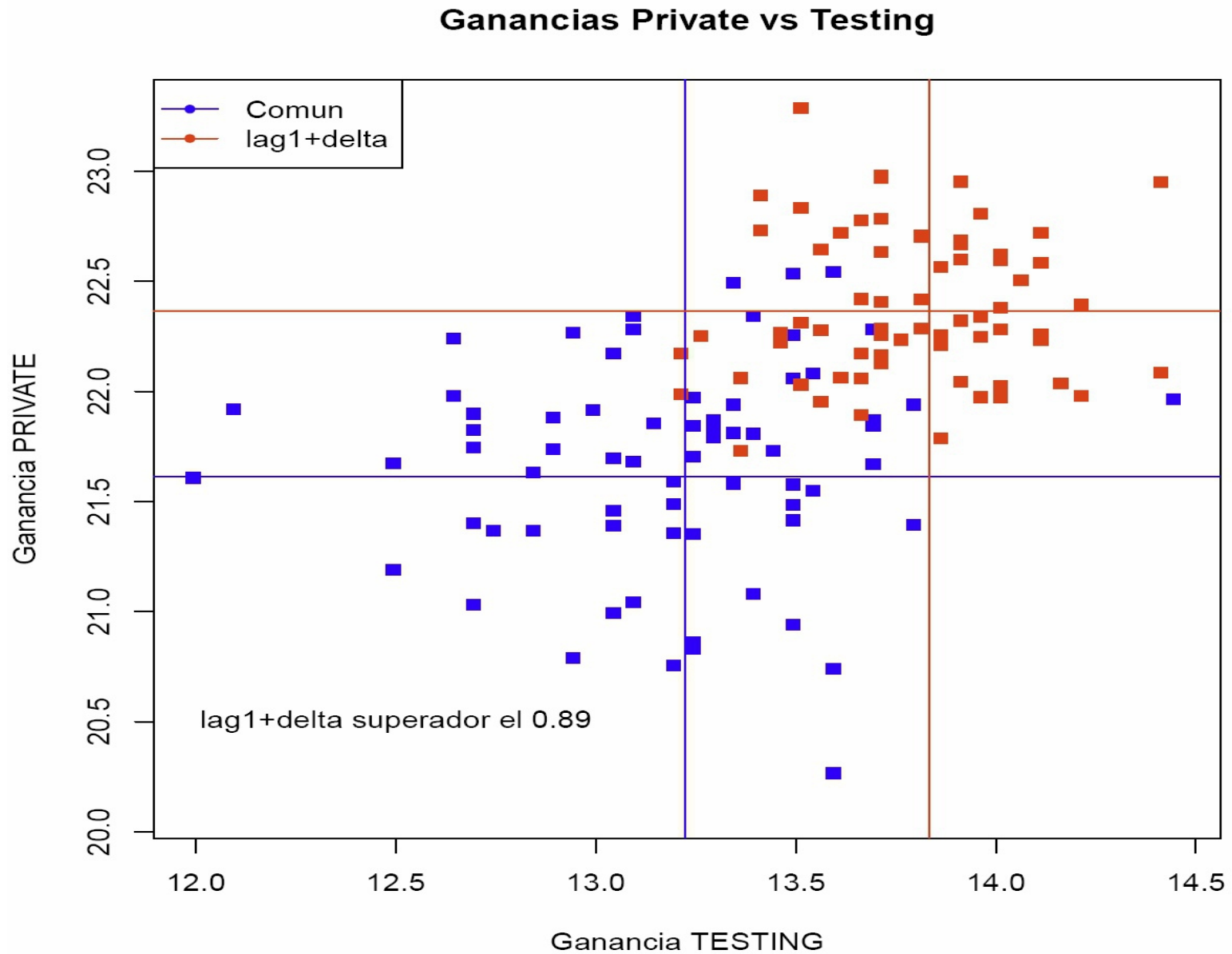
Finalmente, la comparación entre los
dos experimentos

Experimento 2 vs 1

Ganancias Private vs Public



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 [931_wilcoxtest.r](#)