

Ensamble de modelos boosting y clustering: Una propuesta para la reducción de la rotación laboral en Uruguay *#

JUNIO, 2015

**Ec. Oriana Aires
Ec. Santiago Escobar
Msc. Ec. Mauricio Giacometti
Ec. Martin Pereira
Ec. Paulina Rochón
Ec. Micaela Suárez**

Resumen

Es bien sabido que en contextos de baja tasa de desempleo, la rotación del personal se transforma en un problema para la mayoría de las empresas independientemente de la región geográfica en donde nos encontremos.

En el desarrollo del presente trabajo se utilizaron diversas metodologías con el objetivo de maximizar la utilidad esperada de retener a un empleado mediante incentivos económicos utilizando técnicas de *boosted regression* tanto con distribución logística como con distribución normal, así como clusters para lograr los resultados deseados.

El modelo principal es aquel donde la variable dependiente es binaria e indica como éxito que el individuo esté buscando sustituir su trabajo actual. Una de las variables que detectamos como importantes para explicar esa variable de interés es la diferencia entre el salario percibido realmente y el salario esperado que recibiría una persona con su mismo perfil. Para ello se estima un modelo lineal donde la variable dependiente es el logaritmo del salario líquido percibido por hora trabajada. Los errores de esta regresión se incorporan de manera específica en la regresión principal. En ambas regresiones, hay información relevante que está presente de manera cualitativa y para ello se realizó una reducción de dimensiones mediante la utilización de clustering k-means.

Los resultados obtenidos demuestran de manera contundente que cada una de las regresiones estimadas aprueban correctamente los tests usuales de validación y ofrecen altos valores de los indicadores de bondad de ajuste.

Clasificación JEL: C52; C61; J2; M51.

Palabras clave: Mercado laboral uruguayo; rotación laboral; boosted; cluster; k-means.

1. INTRODUCCIÓN

Es sabido que existen ciertas circunstancias que pueden inducir a un trabajador dependiente a querer sustituir su empleo actual por uno nuevo. La alta rotación en los puestos de trabajo en nuestro país, debido a diferentes motivos, es una problemática para aquellas empresas en donde no es fácil aplicar una metodología eficaz de retención. A su vez los procesos de selección son cada vez más complejos y largos implicando un incremento de los costos asociados a estos últimos. Por otra parte, las empresas se están especializando cada vez más, por lo que los períodos de capacitación en tareas específicas son más largos y onerosos ya que la baja del salario al inicio del período laboral no es suficiente para compensar la baja productividad inicial de cada trabajador. Esto se ve intensificado en aquellos puestos de trabajo con tareas más especializadas.

La presente investigación tiene la intención de elaborar un modelo capaz de predecir con el mayor nivel de precisión, si un trabajador está buscando sustituir su trabajo actual por uno nuevo, en base a la información sobre las condiciones de la relación de dependencia, y otras que son intrínsecas de cada individuo.

Para ello, fueron combinadas varias técnicas de distintas disciplinas en lo que respecta al tratamiento de Datos: La técnica de *Boosting* (*Boosted Normal & Boosted Logistic*) y la técnica de segmentación *k-means*.

El modelo resultante consiste en un ensamble de varios modelos y técnicas propias de econometría y estadística. Los resultados obtenidos de esta combinación de modelos se vinculan con una función de utilidad que influye sobre la decisión del empleador de actuar sobre aquellos individuos con alta probabilidad de querer cambiar su actividad laboral, en el entendido de que tanto la decisión de actuar o no hacerlo presenta costos asociados.

Nuestro objetivo principal no se basa en la interpretación económica de la influencia de las diferentes variables en el resultado final de querer sustituir su trabajo en cuanto a signo y magnitud del efecto parcial, sino que se utilizaron todos los medios disponibles, tanto técnicos como de información, para lograr predecir de la mejor manera posible la probabilidad de que un individuo tenga intención de cambiar de trabajo. Esta información será insumo para el departamento de Recursos Humanos de una empresa, de manera tal que la elección de incentivos para la retención de personal así como para contratar nuevo personal, sea realizada de manera óptima. Esto quiere decir que el modelo debe tener tanto un buen poder de asignación ordinal como cardinal ya que el punto de corte es el que define en última instancia la decisión de los analistas de recursos humanos.

Trabajamos con información de individuos que trabajan en empresas medianas y grandes, entendiendo que éstas son las que cuentan con la información y los procesos necesarios para poder aplicar este tipo de complejidad a sus métodos de selección y retención, así como también poseen cierta "cultura de retención", en donde los procesos de selección y de retención son centrales para la explicación de la continuidad del personal asociado.

2. DESCRIPCIÓN DE LA MUESTRA

El análisis se realiza a partir de la Encuesta Continua de Hogares (ECH) para el año 2013 realizado por el Instituto Nacional de Estadística de Uruguay (INE).

La muestra de estudio fue acotada con el fin de lograr una muestra que se ajuste a nuestras necesidades. Debido al objetivo del trabajo, solo nos interesan las personas que estén ocupadas, esto es que forme parte de la población en edad de trabajar (PET)¹ y que efectivamente trabajen de forma remunerada.

A su vez utilizamos solo las personas que residan en Montevideo, dado que tenemos más información para este departamento y la mayoría de las empresas dispuestas a utilizar esta metodología se encuentran en la capital del país. Entendemos que los ratios de la rotación laboral y los motivos de la misma varían considerablemente entre las empresas del Interior y Montevideo y que la incorporación de información de trabajadores del interior del país incorporaría un cambio estructural al modelo que generaría más costos en términos de complejidad y tiempo de análisis que beneficios y por lo tanto dejaremos ese análisis para un trabajo posterior.

Por último, respecto a la relación del trabajador con la empresa y tal como se mencionó en la introducción, nos focalizamos en los trabajadores dependientes remunerados que trabajen en medianas y grandes empresas². Esto se debe a que dado nuestro objetivo, no consideramos apropiado tomar en cuenta los cuentapropistas y patrones ya que no necesitan de un modelo para saber si ellos mismos desean cambiar de trabajo. El fundamento para truncar la muestra en trabajadores que se desempeñan en empresas de mediano a gran porte se debe a que entendemos que en empresas muy chicas, la relación empleador-empleado suele ser informal siendo común que existan lazos familiares o de amistad entre ellos, no existiendo incentivos claros para la rotación laboral.

Respecto a las variables utilizadas en el modelo, se incorporaron variables relacionadas al individuo y al trabajo del mismo. Respecto al individuo se utilizaron: edad, educación, género, barrio en el que vive y horas que trabaja en el hogar. Mientras que relacionadas al trabajo: ingreso, medio de transporte utilizado para ir a trabajar, cantidad de trabajos, si aporta a alguna caja de jubilaciones, si cobra aguinaldo, años de trabajo actual, tamaño de la empresa, si trabaja en un local de la empresa, si estuvo desocupado en los últimos doce meses y a que se dedica la empresa en la que trabaja.

En el anexo del presente trabajo se podrá observar en detalle las variables utilizadas y las estadísticas descriptivas más relevantes de las mismas.

¹ PET: Población en edad de trabajar (Personas mayores a 14 años).

² Se considera mediana y grandes empresas aquellas con al menos diez empleados.

3. METODOLOGÍA

3.1 Algoritmo Boosted

En lugar de utilizar los modelos de regresión tradicionales en Econometría, aplicamos una técnica que se desarrolló hace relativamente poco tiempo para Stata, el algoritmo de *Boosting*³. Este último logra mejorar considerablemente la performance del modelo en cuanto a los indicadores de bondad de ajuste y con ello sus predicciones.

Este algoritmo es una técnica popular en los análisis de tipo *Data mining*. No se basa en la teoría económica para la inclusión de las variables explicativas sino que se incluyen todos los posibles regresores y el modelo distingue aquellas variables relevantes y cuales no lo son, no interfiriendo negativamente estas últimas en los resultados ni en el ajuste del modelo. Una vez terminada la etapa de estimación, se realiza la validación del modelo.

En sus orígenes el algoritmo *boosting* fue creado por Ingenieros en Sistemas para problemas binarios y luego fue adaptado por Estadísticos con el fin de utilizarlo para crear modelos de regresiones secuenciales para variables dependientes con otras características⁴.

El algoritmo automáticamente divide los datos entre la muestra de entrenamiento y de *testing*, a diferencia de otros modelos en los que depende del analista hacer énfasis en el testeo y validación del modelo⁵. Tal como el resto de los modelos, se ajusta en la muestra de entrenamiento, utilizando el resultado para realizar las predicciones en la muestra de *testing*.

Influencia de las variables

Mientras que en los modelos de regresión lineal el efecto de las variables independientes sobre la dependiente es observado a partir de los valores de sus coeficientes (β), el modelo *boosting* introduce el concepto de influencia de las variables. Es decir, como resultado indica el peso (porcentaje) que tiene cada variable para explicar la variable dependiente. A su vez, la influencia no es sensible a la unidad de medida.

Como se mencionó al introducir el algoritmo, el mismo logra aumentar de forma considerable la performance del modelo. M. Schonlau (2005) a través de un ejemplo

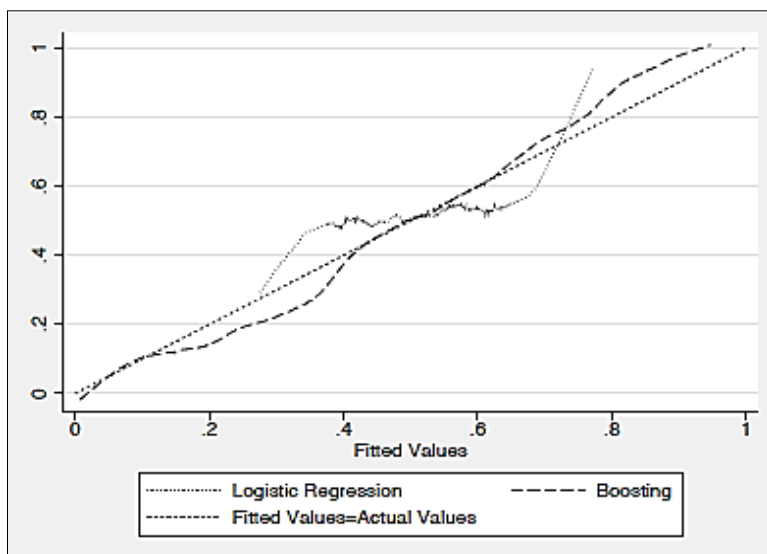
³ La funcionalidad para el programa Stata fue desarrollada por Matthias Schonlau en base al algoritmo *boosting* de gradiente de Friedman.

⁴ En el anexo del presente trabajo puede profundizarse sobre la metodología del algoritmo *boosting* de gradiente de Friedman.

⁵ Por defecto el algoritmo *boosting* separa el 80% de los datos para entrenamiento y el 20% restante para *testing*. De todas formas, mediante el parámetro *trainfraction(#)*, esto puede ser ajustado por el analista.

muestra la mejora de ajuste aplicando *boosting* a una regresión logística. En cuanto a la clasificación a la hora de validar el modelo, con un modelo logístico un 52% de los datos están clasificados correctamente mientras que aplicando el algoritmo se alcanza un 76%. Respecto al *Pseudo-R²*, para el modelo *logit* es de 0,02 mientras que con el *boosting* asciende a 0,27.

En el siguiente gráfico podemos observar la mejora en el ajuste del modelo al utilizar *boosting* en lugar de una regresión logística.



Fuente: The Stata Journal (2005) 5, Number 3, pp. 330–354. Boosted regression (boosting): An introductory tutorial and a Stata plugin Matthias Schonlau.

M. Sconlau (2005) plantea algunos lineamientos acerca de cuándo es conveniente utilizar *boosting*. Tal como se observa en la siguiente tabla, hay algunas circunstancias en donde es preferible la utilización de este algoritmo. Entre ellos, la existencia de muchas variables explicativas, pudiendo darse el caso de tener más variables que observaciones, la utilización de variables interactivas y categóricas, modelos no lineales y también cuando existe una correlación muy alta entre variables.

Indicator	Indicator favors the use of boosting	Indicator against the use of boosting	Why?
small dataset		x	linear approximation usually adequate
large dataset	x		nonlinearities and interactions likely
more variables than observations (or close)	x		linear (Gaussian and logistic) regression methods fail
suspected nonlinearities	x		nonlinearities need not be explicitly modeled
suspected interactions	x		interactions need not be explicitly specified
ordered categorical x-variables	x		awkward in parametric regression
correlated data	x		potential for overfitting
x-variables consist of indicator variables only		x	nonlinearities cannot arise from indicator variables; interactions still might

Fuente: The Stata Journal (2005) 5, Number 3, pp. 330–354. Boosted regression (boosting): An introductory tutorial and a Stata plugin Matthias Schonlau

El comando boosting

Sintaxis:

```
boost varlist [if ] [in] , distribution(string) maxiter(#) [influence predict(varname)  
shrink(#) bag(#) trainfraction(#) interaction(#) seed(#) ]
```

Opciones:

- **Distribution:** especifica la distribución del término de error. Posibles distribuciones pueden ser normal, logística y poisson.
- **Maxiter:** especifica el máximo número de árboles a ser ajustados. El número real usado, *bestiter*, es obtenido como output.
- **Influence:** muestra el porcentaje de variación explicado por cada variable independiente.
- **Predict (varname):** predice y guarda la predicción en la variable *varname*.
- **Shrink:** es el tamaño de la actualización que aplica de iteración a iteración. Afecta positivamente el coeficiente de ajuste pero incrementa drásticamente el tiempo de ejecución del comando.
- **Bag:** especifica la fracción de las observaciones de entrenamiento usadas para ajustar un árbol individual.
- **Trainfraction:** especifica el porcentaje de datos para usar como datos de entrenamiento.
- **Interaction:** máximo de interacciones permitidas.
- **Seed:** especifica la semilla aleatoria para generar la misma secuencia de números aleatorios. El valor por defecto es *seed (0)*.

Funcionamiento

El comando *Boost* determina el número de iteraciones que maximizan la verosimilitud o, lo que es equivalente, el *Pseudo- R^2* . Al contrario del R^2 que proporciona el comando *regress* (utilizado para modelos lineales), el *Pseudo- R^2* en este caso es un estadístico calculado con valores fuera de la muestra.

Outputs y valores de salida

El comando, al finalizar de estimar el modelo, da como resultado un output donde muestra los siguientes valores: mejor número de iteraciones (*bestiter*), el R^2 computado en la muestra de entrenamiento, R^2 computado en la muestra de *testing*, el número de observaciones usado para la muestra de entrenamiento y la influencia de cada variable independiente sobre la dependiente.

3.2 Construcción de Clusters

Es muy común que en los modelos econométricos se utilicen variables cualitativas como regresores. A veces éstas cuentan con un gran número de categorías, lo que implica perder muchos grados de libertad si se decidiera introducir en el modelo cada una por separado (en forma de *dummies*).

Dado que en nuestro set de datos disponemos de varias variables cualitativas que constan de muchas categorías, decidimos incorporar dicha información a través de clusters en vez de utilizar variables *dummies* ya que de esta manera el modelo quedaría poco parsimonioso. De esta forma, se logra una reducción de dimensiones con la menor pérdida de información posible. Dichos clusters se construyen de forma tal que las categorías sean lo más similares posibles dentro de los grupos y disimilares con el resto.

Los clusters generados se utilizaron para incorporar como variables independientes al modelo logit que actúa como regresión principal y para el modelo lineal, que actúa como regresión auxiliar. Dicha información fue incorporada en forma de *dummies*.

Las variables elegidas para clusterizar fueron las siguientes:

- Características de la empresa en donde los individuos trabajan, caracterizadas por el agrupamiento CIIU⁶ (122 categorías de actividad).
- Ubicación geográfica de la vivienda de los individuos, dividida por barrios de la capital del país (62 barrios).
- Situación educativa, en donde se determina la trayectoria del individuo durante el ciclo académico (510 combinaciones).

Respecto a la variable situación educativa, ésta fue construida con el fin de poder captar la trayectoria educativa de los individuos. Tomando en cuenta primaria, secundaria, terciaria y posgrado. Asimismo, se diferencia si se completó cada instancia y si cada una de estas es privada o pública.

A modo de ejemplo, la variable situación educativa de un individuo muestra lo siguiente:

Nivel	Tipo	Finalización
Primaria	Pública	Si
Secundaria-ciclo básico	Pública	Si
Secundaria-bachillerato	Privada	Si
Universidad	Pública	Si
Postgrado	Pública	No

Para cada variable cualitativa, se realizaron dos clusters, uno para cada una de las regresiones estimadas. Los mismos cuentan con dos variables de agrupación, una de

⁶ Clasificación Industrial Internacional Uniforme. www.ine.gub.uy/biblioteca/ciiu4/estructura%20ciiu4.pdf

ellas es la cantidad de individuos por categoría de la variable cualitativa y otra en base a la media de la variable dependiente. Esto se realizó en el entendido que es deseable que las categorías con pocos individuos se agrupen de manera tal que tengan efecto real sobre la performance del modelo y además deben ser similares en cuanto al efecto que tengan sobre la variable dependiente de cada modelo.

El único parámetro que requiere el algoritmo *k-means* es la cantidad de grupos (k) que se quiere construir.

Se realizaron todos los clusters con diferentes k y posteriormente se realizó un análisis, quedándonos con los que tienen mejor performance. Dicho análisis consistió en realizar una regresión solamente con las categorías como regresores y observar un indicador de performance (R^2 corregido en el caso de la regresión lineal asociada al ingreso y *Akaike info criterion* –AIC- en el modelo asociado a la decisión de cambiar de trabajo). De esta forma, se observa cual es la cantidad de grupos (k) óptima para cada cluster.

3.3 Ensamblaje del modelo

El modelo de interés para obtener la probabilidad estimada de que un individuo esté buscando trabajo, es un *logit* en donde la variable dependiente es binaria, tomando el valor 1 si la persona está buscando cambiar de trabajo (YBT) y 0 en otro caso. Este modelo tiene variables explicativas que fueron mencionadas en la sección anterior, y también los errores de una regresión auxiliar incorporados de una manera particular.

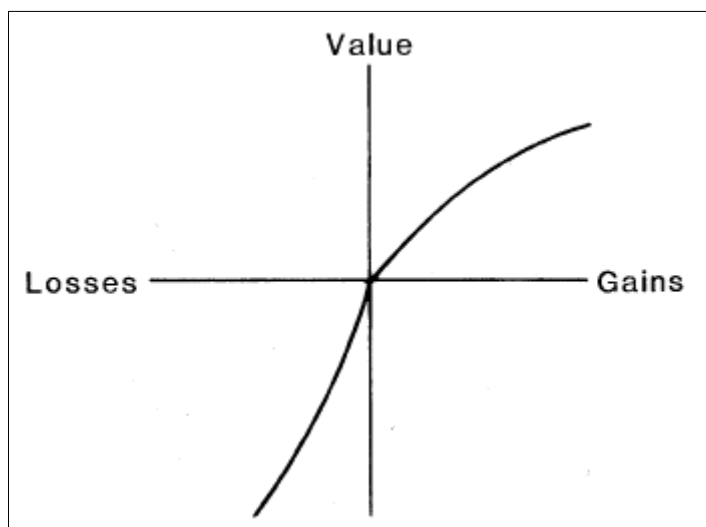
Los errores de la regresión auxiliar actúan como una variable independiente más de la regresión principal. Lo que se intenta medir es la diferencia entre el salario que está percibiendo el individuo y el salario que el mercado está efectivamente pagando en promedio por un empleado con su mismo perfil.

Dado que nuestro objetivo final es medir la probabilidad de que un trabajador esté buscando empleo para sustituir el actual, dicha variable pasa a ser vital en la predicción de este comportamiento.

La intuición es que si un individuo está percibiendo menos de lo que el mercado está pagando en promedio por alguien con sus características, tendrá incentivo a cambiar su trabajo. En cambio aquellas personas que están ganando por encima de lo que el mercado laboral remunera en promedio serían menos propensas a cambiar de trabajo.

Para incorporar esta variable en la regresión principal se tomó la decisión de crear 2 variables con la información obtenida previamente, construyéndolas de la siguiente manera: **uhatpos** toma el valor de la diferencia entre el salario percibido y el salario de mercado cuando para un individuo con su mismo perfil es positiva y cero en otro caso; y la variable **uhatneg** sigue la misma lógica pero si la diferencia observada es negativa y cero en otro caso, de manera que el producto interno es cero. Esto se realiza en el entendido de que existe un comportamiento diferenciado para quienes están en el dominio negativo de la misma con respecto a los que están en el dominio positivo.

Esta división se basó en la propuesta teórica denominada *Prospect Theory*⁷ en donde se demuestra que el comportamiento de los individuos presenta una función convexa para las pérdidas y una función cóncava para las ganancias.



Fuente: Amos Tversky and Daniel Kahneman, 1981. The Framing of Decisions and the Psychology of Choice. Science, New Series, Vol. 211, No. 4481. (Jan. 30, 1981), pp. 454.

La imagen muestra como la motivación de las personas es mayor en un ambiente de pérdidas que en un ambiente de ganancias. De hecho se puede observar que la motivación no se incrementa indefinidamente con los beneficios que representa sino que se estanca. Sin embargo, el miedo decrece rápidamente al principio y luego lentamente a medida que las pérdidas crecen. Esto representa una evaluación diferente en la toma de decisiones en el dominio de las ganancias que en el de las pérdidas.

Los individuos valoran en mayor medida cambiar de trabajo cuando se encuentran percibiendo ingresos por debajo de lo que el mercado está dispuesto a pagar por su perfil. A diferencia de aquel individuo, uno que se encuentra en la situación contraria, a priori tendría menos incentivos.

Una vez generada esta división, hemos de esperar que una vez incorporado tanto al modelo *logit* como al *boosted logistic*, estas variables tengan un comportamiento distinto entre sí y que además, los individuos que están por debajo del salario de mercado (*uhatneg*) tenga mayor porcentaje de explicación que *uhatpos* que representa a los individuos que están por encima de lo que les pagaría en promedio otra empresa en el mercado.

A su vez, el ingreso fue determinado como el ingreso que percibe el individuo por su ocupación principal en relación de dependencia. A este agregado le descontamos los ingresos por aguinaldo y salario vacacional, con la finalidad de evitar perturbaciones derivadas del momento del año en que se declaran los ingresos por parte del encuestado.

⁷ Kahneman, D. y Tversky, A. (1981). "The Framing of Decisions and the Psychology of Choice". Por apuestas simples, los sujetos tienden a evitar riesgos en el dominio de las ganancias y tienden a buscar los riesgos en el ámbito de las pérdidas. (Pp. 2)

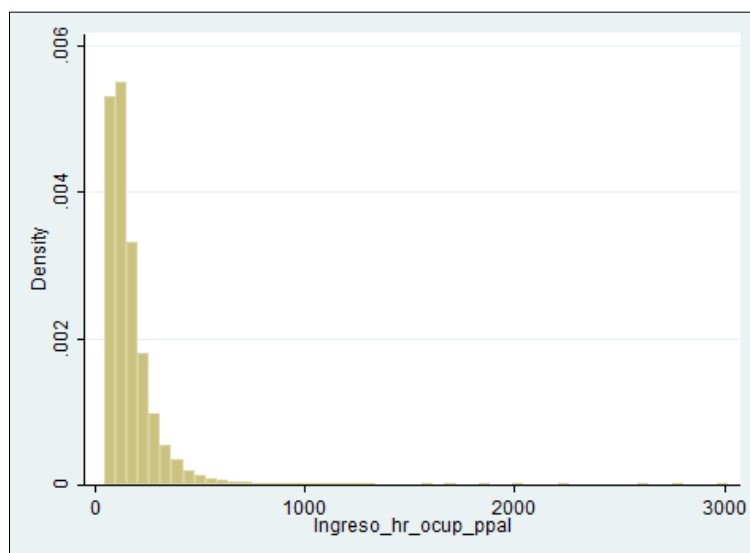
A este último agregado lo llevamos de una unidad de medida semanal, a una unidad de medida por hora, siendo la variable resultante el ingreso de un individuo por hora obtenido de su ocupación principal.

Por no poseer una distribución normal, el salario hora es transformado en una función logarítmica dado que estaría mejor comportado hacia una distribución más aproximada a normal. Con esta corrección es posible obtener los porcentajes de los coeficientes estimados en las regresiones multivariadas.

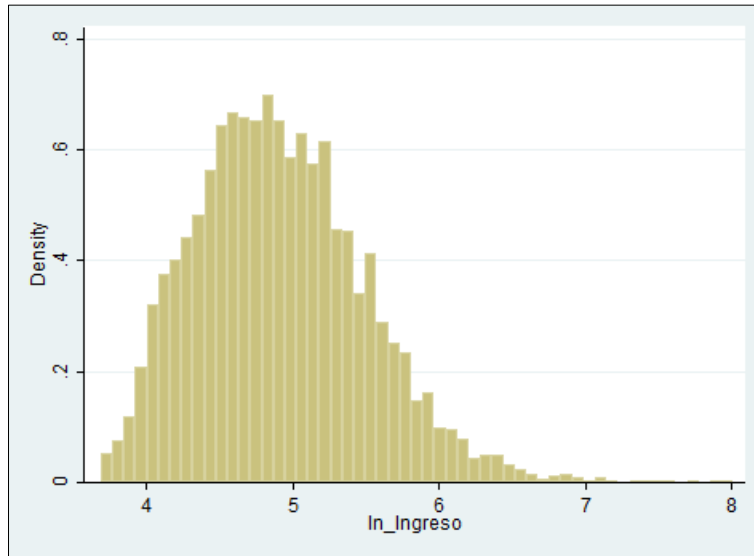
Durante el proceso de revisión de la calidad de datos, encontramos una nueva causalidad de individuos en donde, basados en el histograma, pudimos verificar que existían observaciones concentradas en un salario muy bajo, y que resultaban siendo *outliers* en cualquier especificación estimada del modelo lineal.

Al realizar la evaluación de estos individuos y comparándolos con el salario mínimo nacional concluimos que era eficaz excluir dichos *outliers*, ya que no eran representativos en la muestra y hacían que ésta tomara valores atípicos del salario del 2013. Para esto, excluimos aquellas observaciones en donde el salario por hora era menor a \$ 40, resultando ser en el entorno de las 20.000 observaciones (alrededor del 6%).

En los siguientes gráficos se pueden observar un histograma del logaritmo del ingreso por hora y del logaritmo del ingreso:

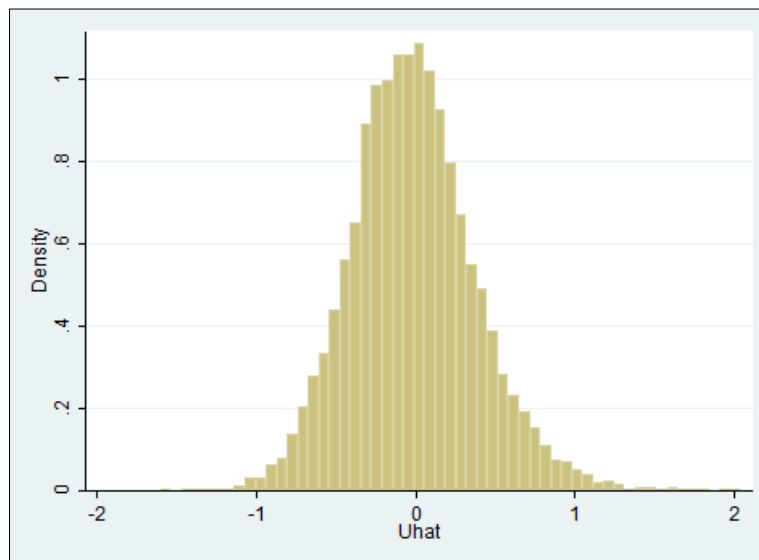


Fuente: Elaboración Propia



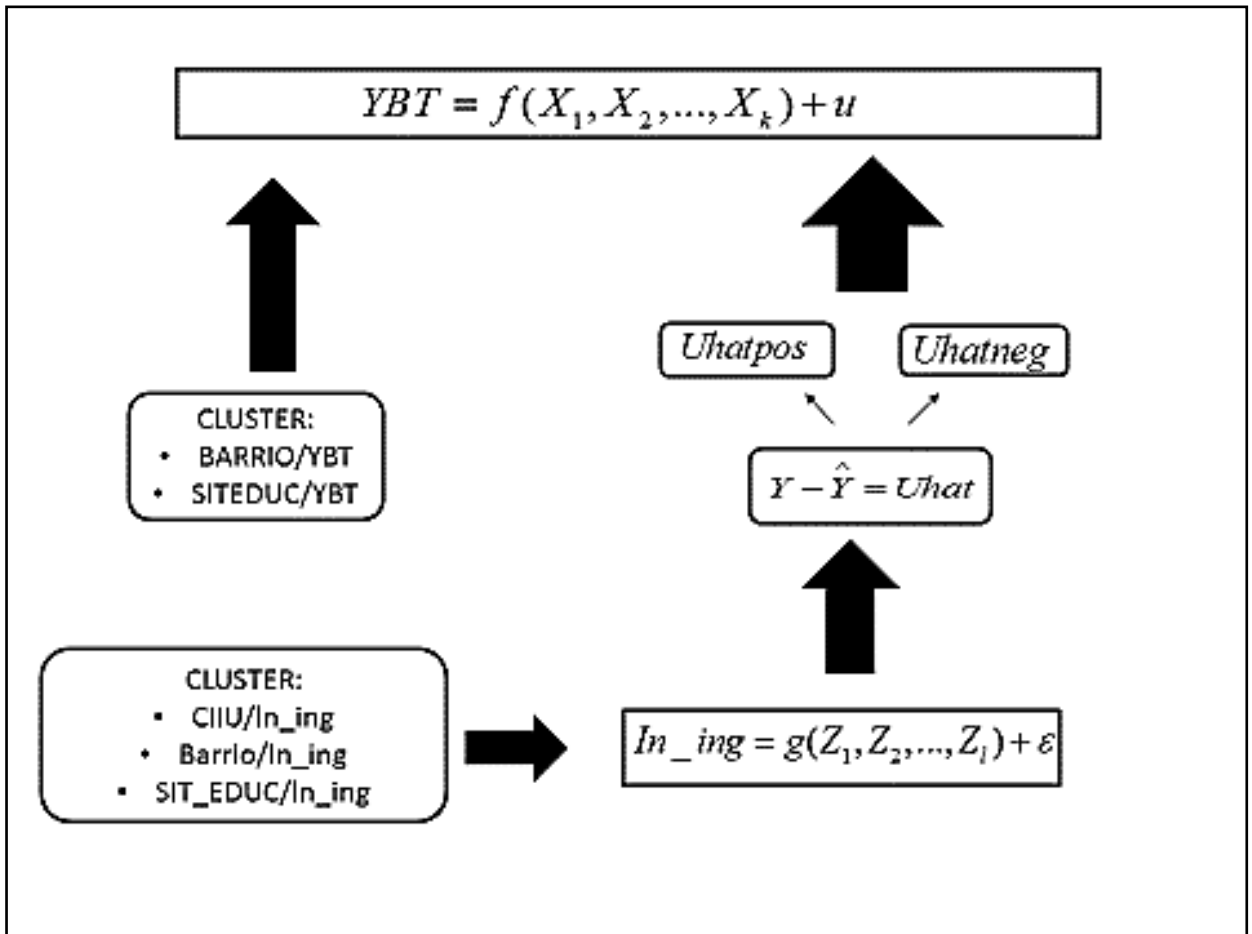
Fuente: Elaboración Propia

En el siguiente gráfico se puede observar la distribución de los errores de la regresión auxiliar:



Fuente: Elaboración Propia

Representación gráfica del ensamblaje del modelo



Fuente: Elaboración Propia

Mediante el diagrama intentamos transmitir como los flujos que se generan desde los clusters se vinculan con las regresiones.

Utilizamos en nuestro primer paso, 3 clusters para la regresión auxiliar en donde toman la forma de regresores (z).

Luego de creada la regresión auxiliar, de ahora en adelante $g(z)$, extrajimos el error para poder trabajar en nuestra regresión madre con él como variable independiente (x).

Respecto a la estimación de los modelos, la regresión principal se modeliza mediante un *logit*, el cual se estima utilizando el algoritmo *Boosted Logistic*.

En cambio, la regresión auxiliar, es un modelo lineal, estimándose a través del algoritmo *Boosted Normal*.

4. RESULTADOS

4.1 Resultados para la regresión auxiliar

En la siguiente tabla se puede observar el resultado del modelo boosted para la regresión auxiliar⁸.

Variable	Influencia
Educación	38,52
Años en el trabajo actual	21,37
Edad	19,41
Horas de Trabajo en el Hogar	3,21
Género	2,8
Cantidad de trabajos	2,31
Tamaño empresa (entre 10 y 19 empleados)	1,55
Cluster educación	1,28
Cluster barrio	1,1
Tamaño empresa (entre 20 y 49 empleados)	1,03

Fuente: Elaboración Propia

A la luz de los resultados obtenidos, la variable con mayor peso resulto ser la cantidad de años de educación formal terminados. Como era de esperar, la educación es un variable determinante basados en los enfoques de capital humano y la relevancia cada vez mayor que tiene esto sobre el mercado laboral uruguayo.

En segundo lugar en términos de influencia, está la variable que representa los años de trabajo en el trabajo actual. Esto es consecuente con la política de negociación salarial propuesta por el gobierno a partir del 2005 basada en los consejos de salarios.

La edad influye positivamente sobre el logaritmo del salario por hora debido a la relación que hay entre ésta y la experiencia que se genera con el pasar de los años.

La cantidad de horas de trabajo que el individuo realiza en su hogar es también una de las que más influye, ya que expresa mayor cantidad de horas trabajadas y mayor nivel de responsabilidad en la tarea realizada.

La variable género que es una *dummy* que toma valor 1 si el individuo es mujer, es también una variable relevante para nuestra regresión, lo que estaría relacionado con la brecha de salario entre géneros evidenciada en muchos trabajos centrados en el mercado laboral nacional.

⁸ Solo se presentan las variables que influyen al menos en un 1%

La cantidad de trabajos que el individuo tiene al mismo tiempo, es también una variable de relevancia para la explicación del salario de la ocupación principal.

Si la persona pertenece a una empresa de tamaño chico (empresas de entre 10 y 19 empleados) influye más que si trabaja en una empresa de tamaño medio (tiene entre 20 a 49 empleados).

4.2 Resultados para la regresión principal

En la siguiente tabla se puede observar la influencia que tienen las variables explicativas sobre la decisión de buscar otro trabajo.

Variable	Influencia
Uhatneg	33,74
Edad	17,92
Años en el trabajo actual	11,52
Uhatpos	10,13
Educación	6,79
Género	3,75
Cluster barrio	1,89
Tamaño empresa (más de 50 empleados)	1,12
Desocupado (últimos 12 meses)	1,07
Cluster barrio	1,02
Cantidad de trabajos	0,99
Cluster barrio	0,94
Cluster barrio	0,94
Cluster barrio	0,91
Cluster educación	0,89
Cluster educación	0,87
Cluster educación	0,83
Tamaño empresa (entre 10 y 19 empleados)	0,82
Cluster educación	0,77
Cluster barrio	0,71
Cluster educación	0,64
Horas de Trabajo en el Hogar	0,63
Aporta a caja de jubilaciones	0,49
Tamaño empresa (entre 20 y 49 empleados)	0,39
Cobra aguinaldo	0,17

Fuente: Elaboración Propia

Tal como era de esperar, la variable con mayor influencia en que un individuo esté buscando cambiar de trabajo es uhatneg, esta variable representa la diferencia entre el

salario percibido y el salario de mercado de cada individuo (siendo positiva cuando dicha diferencia es mayor a cero). Por lo tanto, se comprueba que los individuos tienen fuertes incentivos a la rotación laboral cuando están cobrando un salario por debajo del salario de mercado.

A su vez, la edad y años de trabajo en el trabajo actual influyen de forma importante en que un individuo busque cambiar de trabajo, tal como lo hacen en el logaritmo del salario (regresión auxiliar).

La variable *uhatpos* tiene una influencia considerable en la variable dependiente, siendo este un resultado no muy trivial ya que esta variable representa la diferencia negativa entre salario de mercado y el salario del individuo, esto es que se percibe un salario mayor al salario del mercado.

Las variables educación, género y cantidad de trabajos muestran ser importantes también a la hora de tomar la decisión de buscar cambiar de trabajo.

Por último, se observa una incidencia positiva y significativa de los diferentes cluster de educación, barrio y tamaño empresa⁹.

4.3 Elección del punto de corte

El punto de corte es utilizado como probabilidad umbral a partir de la cual se toma la decisión de actuar sobre los individuos detectados con alta probabilidad de sustituir su trabajo. Visto de otra forma, es el que define que una observación sea considerada con un \hat{Y} igual a 1 o 0.

Es por esto que se entiende que es un tema de conveniencia de la empresa y se transforma en una elección que debe tomarse desde una óptica económica y no estadística. Se debe elegir de manera tal que incremente el valor de la decisión en la empresa y no como el punto que mejor discrimina a los individuos.

Si elegimos el punto de corte muy bajo y los costes asociados a actuar sobre el problema para prevenirlo son relativamente altos, será peor el remedio que la enfermedad, ya que actuaremos sobre demasiados sujetos generando costos que se elevarían por encima de los beneficios. Un caso similar pero opuesto sucede si elegimos el punto de corte demasiado alto, solamente actuaremos sobre pocos individuos y hubiese sido rentable actuar sobre más, dado que los costos lo ameritaban. Es por lo tanto un asunto de elección del punto que maximice la utilidad esperada ($E(U)$) en función de los esfuerzos que cada empresa esté dispuesta a realizar para retener a cada persona de su *staff*.

Es por todo lo anterior que para la elección del mismo se utiliza una función de utilidad y se definirán como probabilidades altas aquellas para las cuales el valor esperado de la utilidad resulte positivo.

⁹ El detalle de la influencia de cada grupo de los clusters está a disposición por pedido.

La función esperada de utilidad es del tipo $E(U) = X(1-p) - Zp$ siendo X la ganancia obtenida dado que la persona no esté buscando sustituir su trabajo actual, Z la pérdida dado que la persona decide abandonar su puesto de trabajo y P la probabilidad de que la persona quiera sustituir su trabajo actual.

Para poder realizar el análisis y poder darle una mejor interpretación a los resultados, es que tomaremos ciertos supuestos acerca del comportamiento del individuo frente a su trabajo. Tomaremos la situación de estar buscando trabajo como que realmente lo sustituye, y en el caso de no estar buscando sustituir su trabajo que la persona no se va de su lugar de trabajo. En una economía como la uruguaya en estos momentos, donde el desempleo se encuentra en niveles históricamente bajos, el hecho de estar buscando sustituir un trabajo por otro se puede leer perfectamente como que la persona realmente sustituye su trabajo actual por un nuevo trabajo. A su vez, el no estar buscando sustituir el trabajo actual se puede entender como que la persona se quedará en su trabajo actual, ya que la búsqueda de trabajo es una condición necesaria para el cambio efectivo.

Para la elección de las variables de costos e ingresos de tomar la decisión de retener al empleado (X y Z) realizaremos los siguientes supuestos a los efectos de poder monetizar la información que nos arroje el modelo principal de este trabajo. Con respecto a X debemos analizar que costo nos ahorraremos (ganancia) si el individuo decide quedarse en su actual trabajo.

En este sentido encontramos ahorro en costos causados por la ineficiencia en la realización de las tareas por personas que hasta ese momento no habían realizado nunca las acciones de la persona que decidió cambiar su trabajo, a esto se le suma que dejaran de dedicarle todo su tiempo a sus tareas para cubrir a la persona que se fue de la empresa. Otro costo que ahorraremos es el de buscar a otra persona para cubrir el puesto de trabajo vacante y por último mencionaremos que una vez encontrada la persona que se cree indicada para el puesto, se necesitará un período de adaptación y aprendizaje, por lo que nos encontramos frente a un nuevo costo a ahorrar, asociado a la curva de aprendizaje.

Con respecto a los costos en los cuales deberá incurrir la empresa para que la persona no se vaya de la misma, que llamamos Z , cada empresa evaluará entre distintos métodos de incentivos u otras herramientas para que la persona no abandone su puesto de trabajo como por ejemplo: Bonos, promociones, relocalización, nuevas tareas, nuevas responsabilidades, relocalización geográfica, entre otras. Cada uno de ellos tiene diferentes costos asociados, algunos serán costos de oportunidad, otros serán costos financieros, de mayor o menor porte.

Entonces, partiendo de igualar nuestra función de Utilidad a cero obtenemos lo siguiente:

$$X(1-p) - Zp = 0 \Rightarrow X(1-p) = Zp$$

Y desarrollando podemos llegar al valor p para la muestra:

$$\frac{Z+X}{X} = \frac{1}{p} \Rightarrow p = \frac{X}{Z+X}$$

5. EVALUACIÓN MONETARIA

En general, los estudios que implican estimaciones de modelos para variable dependiente binaria, terminan calculando indicadores de proporción de 1's correctamente predichos y de 0's correctamente predichos y es común encontrar modelos que funcionan muy bien prediciendo los 1's, pero no tan bien los ceros y viceversa. Por lo que es difícil determinar el modelo que mejor ajusta al proceso generador de los datos.

En este sentido lo que implementamos fue una valoración objetiva de los resultados en función de ciertos supuestos sobre los costos e ingresos relacionados al abandono del puesto de trabajo. Dichos supuestos se realizan dado que no estamos evaluando este modelo en una empresa puntual, en donde los costos e ingresos son estimables.

La monetización consiste en evaluar en cada caso según la siguiente casuística:

\hat{Y}	Y	U
0	0	0
0	1	$-X$
1	0	$-Z$
1	1	$X - Z$

Es decir que si el modelo detecta que el empleado no va a renunciar y no renuncia, no hay costos ni ingresos asociados a esa situación. Si el modelo aconseja no actuar y el empleado renuncia, la empresa debe cargar con los costos de la rotación. En cambio, si el criterio de decisión indica que el funcionario va a renunciar y el mismo no pensaba tomar esa acción, la organización habrá gastado el costo del incentivo en vano. Por último, si el modelo indica que a esa persona debería otorgársele un incentivo para que continúe en la empresa y la misma estaba pensando en abandonar su puesto de trabajo, se habrán ahorrado los costos de rotación al precio del incentivo.

6. RESULTADOS FINALES

Uno de los Indicadores más revisados, en lo que respecta a la Bondad de Ajuste, de los distintos modelos econométricos es el R^2 (para caso de modelos de variable dependiente binaria sería el *pseudo* R^2). En esta parte del trabajo se quiere mostrar la potencia que tiene la técnica de *Boosted* tanto para regresiones MCO, como para regresiones logísticas, ya que aplicando dicha técnica los valores del R^2 o *pseudo* R^2 mejoran sustancialmente.

En el caso de la regresión "Auxiliar" que se implementó, el R^2 pasó de 0,39 a 0,548, mejorando en casi un 40% el ajuste del modelo MCO.

En el modelo principal, *Logit*, el *pseudo R²* pasó de 0,15 a 0,25, mejorando en gran medida el ajuste del modelo.

El hecho de que las variables generadas por la regresión auxiliar: *Uhatneg* y *Uhatpos* sean altamente influyentes en el modelo principal, es de gran relevancia. La primera es más relevante aún que la segunda, lo cual tiene gran sentido ya que equivale al hecho de que el individuo esté siendo remunerado por debajo de lo que paga el mercado por un sujeto con su mismo perfil.

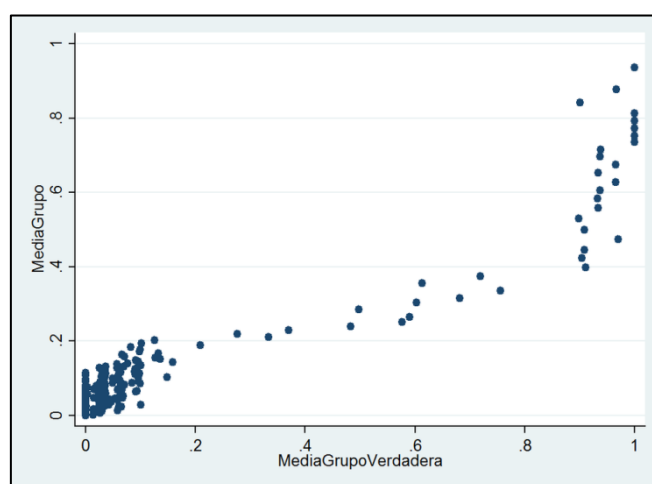
También son considerablemente influyentes la edad, educación y la cantidad de años que el individuo se encuentra empleado en el trabajo actual.

Por otra parte, en el modelo principal se encuentra una diferencia que resulta considerable entre el *R²* de la muestra de entrenamiento y la de *testing*, lo que indica que el modelo está sobre-especificado. Lo anterior es altamente probable en modelos muy flexibles como el resultante del algoritmo *boosted*.

Con el objetivo de observar las consecuencias de la sobre-especificación mencionada anteriormente, se realizó la comparación entre la media de la variable dependiente y la probabilidad estimada.

Se dividieron las observaciones de la variable dependiente estimada en cuatrocientos grupos de igual tamaño. Se calcula la variable *MediaGrupoVerdadera* como la media por grupo de la variable dependiente observada en los datos de la ECH y *MediaGrupo* como la media por grupo de la estimada por el modelo.

El gráfico de dispersión arroja el siguiente resultado:



Fuente: Elaboración Propia

En el escenario ideal, Se espera que el gráfico de dispersión muestre una nube de puntos en torno a una recta de 45° con origen en (0,0). El obtenido a partir del modelo presenta

puntos por debajo de dicha recta, lo que indica que hay una subestimación sistemática.

Sin embargo, hay una alta relación entre el promedio de la variable observada y el de la variable estimada, lo que indica el alto poder ordinal del modelo. Sin embargo tal como lo expresamos anteriormente, hay evidencia de que las predicciones tienden a subestimar la verdadera probabilidad.

Por otra parte, se aprecia que el modelo concentra las observaciones en los extremos del recorrido de la variable real. Por lo tanto, el modelo tiende a discriminar de manera eficiente las observaciones en empleados que quieren sustituir su trabajo y aquellos que no.

La subestimación sistemática de la probabilidad de estar buscando trabajo podría corregirse si se estimara el coeficiente de corrección apropiado.

Una manera de hacerlo es mediante una regresión lineal, en la cual la variable *MediaGrupoVerdadera* sea la dependiente y *MediaGrupo* la independiente. Una vez realizada esta regresión observamos que el coeficiente de ajuste R^2 toma el valor 90%, lo cual reafirma lo observado en la gráfica de dispersión, es decir, que el modelo ordena con extrema precisión la variable de interés.

7. ASPECTOS A MEJORAR

En este apartado se detallan las mejoras a futuro que se consideran relevantes.

Respecto a la implementación de la técnica *Boosting* consideramos muy deseable disponer de una computadora con mayor potencia, ya que la estimación de un modelo mediante este algoritmo en Stata demora mucho tiempo. De esta forma se podrían realizar una mayor cantidad de modelos y así obtener el mejor ajuste posible.

Por ejemplo, sería muy deseable poder trabajar con un *Shrink* menor a 0,1, para esto, según las referencias bibliográficas consultadas se deben aumentar el número de iteraciones en la misma variación en que se modificó el valor de *Shrink*. A modo de ejemplo, si el *Shrink* pasa de 0,1 a 0,01, lo que implica una disminución de 10 veces su valor, debemos aumentar en 10 las iteraciones. Si estas estaban en 2.500, debemos aumentarlas a 25.000. Esto implica un manejo de equipos de mucha potencia para poder trabajar de la mejor manera.

Como se mencionó en el apartado de Análisis Descriptivo, esta investigación está orientada hacia la población de Montevideo. Sin embargo creemos necesario que esta aplicación debe ser ampliada para todo el territorio nacional, incluyendo los departamentos del interior del Uruguay.

Si bien se podría esperar que los comportamientos de los individuos de los otros 18 departamentos sean diferentes, dado el cambio estructural que existe entre Montevideo y el Interior, entendemos que es un aspecto a mejorar de este trabajo.

Otro aspecto a mejorar sería eliminar el supuesto que establecimos en el apartado de la

elección del punto de corte. Este establecía que si un individuo estaba buscando trabajo para sustituir el actual, lo consideraríamos como que abandonaría su puesto de trabajo dadas sus condiciones actuales. Si bien el supuesto es necesario para darle un sentido más útil al trabajo, se debería intentar eliminar de alguna manera para poder realizar la mejor inferencia posible en el lugar que se desee.

Como se mencionó en el trabajo, la distancia entre el R^2 de la muestra de entrenamiento y de *testing* es más amplia que la deseada, lo que indica sobrespecificación, estando asociada a un incremento en la varianza de las predicciones. Por lo cual se plantea como aspecto de mejora, intentar disminuir la brecha existente entre los dos indicadores, mejorando la predicción de los correctos y por lo tanto la aplicación del modelo.

Por último, como se mencionó anteriormente, el modelo está sobreespecificado. De esta forma, al asignar las probabilidades estimadas para cada individuo, se encuentran distorsiones.

Se realizó el cálculo de la media para casi 400 grupos del mismo tamaño, una vez ordenadas las variables de manera descendente en base a la probabilidad estimada. Se compara la media de Y y la media de la probabilidad estimada observándose que hay individuos dentro de la muestra que se encuentran subestimados. Esto genera que, una vez establecido el punto de corte, existan observaciones que deberían tener una mayor probabilidad estimada y por ende algunas tendrán un $\hat{Y} = 0$ cuando deberían tener un $\hat{Y} = 1$.

BIBLIOGRAFÍA

- Adriana Cassoni, (1992). *Una propuesta metodológica para la especificación de modelos econométricos*. Montevideo, Universidad de la República. Facultad de Ciencias Sociales.
- Aris Spanos. (1986). *Statistical foundations of econometric modelling*. Press Syndicate of the University of Cambridge.
- Breiman, L., J. Friedman, R., Olshen, and C. Stone. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- Espino, A. Machado, A. y Alvez, G. (2011). *Estudio de las tendencias cuantitativas y cualitativas de la oferta y la demanda de trabajo en Uruguay: ¿hay un ajuste de la oferta de trabajo a la demanda?* Instituto de Economía de la FCEyA de la UDELAR.
- Friedman, J., T. Hastie, and R. Tibshirani. (2000). *Additive logistic regression: a statistical view of boosting*. *Annals of Statistics* 28: 337-407.
- Galassi, G. y Andrada, M. (2006). *La relación entre educación e ingresos: Ecuaciones de Mincer por regiones geográficas de Argentina para el año 2006*. CEA-CONICET (UNC) – FCE (UNC), CEA-CONICET (UNC) – UNLAR.
- Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. Second Edition. A Wiley-Interscience Publication.
- Long, J. S and Freese, J. (2003). *Regression Models for Categorical Dependent Variables Using Stata*. rev. ed. College Station, TX: Stata Press.
- Matthias Schonlau, (2005). *Boosted regression (boosting: An introductory tutorial and a Stata plugin*. *The Stata Journal*.
- Tversky, A. and Kahneman, D. (1981). *The Framing of Decisions and the Psychology of Choice*. *Science, New Series*, Vol. 211, No. 4481. (Jan. 30, 1981):453-458.

ANEXO

Anexo metodológico

A.1 Presentación de la regresión auxiliar:

$$g(Z_1; \dots; Z_{31})$$

Z_1 = Edad del individuo

Z_3 = Educación (Suma de años completados en la educación formal)

Z_4 = Género (Mujer = 1)

Z_5 = Cantidad de trabajos

Z_6 = Aporta (Si aporta a la caja = 1)

Z_7 = Aguinaldo (Si cobra aguinaldo = 1)

Z_8 = Hrs Trabajo en el hogar

Z_9 = Años trabajo actual

Z_{10} = Tamaño Empresa (entre 10 y 19 empleados)

Z_{11} = Tamaño Empresa (entre 20 y 49 empleados)

Z_{11} = Tamaño Empresa (mas de 50 empleados)

Z_{12} = Local (Trabaja en el local de la empresa = 1)

Z_{13} = Desocupado (Desempleado en los últimos 12 meses = 1)

Z_{14} = Cluster Educación $_{1/7}$

Z_{15} = Cluster Educación $_{2/7}$

Z_{16} = Cluster Educación $_{3/7}$

Z_{17} = Cluster Educación $_{4/7}$

Z_{18} = Cluster Educación $_{5/7}$

Z_{19} = Cluster Educación $_{6/7}$

Z_{20} = Cluster Educación $_{7/7}$

Z_{21} = Cluster Barrio $_{1/7}$

Z_{22} = Cluster Barrio $_{2/7}$

Z_{23} = Cluster Barrio $_{3/7}$

Z_{24} = Cluster Barrio $_{4/7}$

Z_{25} = Cluster Barrio $_{5/7}$

Z_{26} = Cluster Barrio $_{6/7}$

Z_{27} = Cluster Barrio $_{7/7}$

Z_{28} = Cluster CIU $_{1/4}$

Z_{29} = Cluster CIU $_{2/4}$

Z_{30} = Cluster CIU $_{3/4}$

Z_{31} = Cluster CIU $_{4/4}$

A.2 Presentación de la regresión principal:

$$YBT = f(x_1; \dots; X_{25}) + u$$

X_1 = Edad del individuo

X_2 = Educación (Suma de años completados en la educación formal)

X_3 = Uhatneg (dif cuando entre $\ln_ingreso$ y $\hat{\ln_ingreso}$ es positivo, y cero en otro caso)

X_4 = Uhatpos (dif cuando entre $\ln_ingreso$ y $\hat{\ln_ingreso}$ es negativo, y cero en otro caso)

X_5 = Género (Mujer = 1)

X_6 = Cantidad de trabajos

X_7 = Aporta (Si aporta a la caja = 1)

X_8 = Aguinaldo (Si cobra aguinaldo = 1)

X_9 = Hrs Trabajo en el hogar

X_{10} = Años trabajo actual

X_{11} = Desocupado (Desempleado en los últimos 12 meses = 1)

X_{12} = Cluster Educ_{1/5}

X_{13} = Cluster Educ_{2/5}

X_{14} = Cluster Educ_{3/5}

X_{15} = Cluster Educ_{4/5}

X_{16} = Cluster Educ_{5/5}

X_{17} = Cluster Barrio_{1/6}

X_{18} = Cluster Barrio_{2/6}

X_{19} = Cluster Barrio_{3/6}

X_{20} = Cluster Barrio_{4/6}

X_{21} = Cluster Barrio_{5/6}

X_{22} = Cluster Barrio_{6/6}

X_{23} = Tamaño Empresa (entre 10 y 19 empleados)

X_{24} = Tamaño Empresa (entre 20 y 49 empleados)

X_{25} = Tamaño Empresa (mas de 50 empleados)

A.3 Metodología del algoritmo boosting de gradiente de Friedman

El algoritmo boosting utiliza el árbol de regresión como método de clasificación de observaciones.

Friedman, Hastie y Tibshirani (2000) lograron adaptar el método llevándolo a una formulación estadística que utiliza la verosimilitud, creando el algoritmo boosting de regresión logística.

Los pasos que sigue el método son los siguientes:

- establecer estimación inicial como \bar{y} .

Se utiliza como primera aproximación, el valor promedio de la variable dependiente para la predicción de todas las observaciones. Equivale a un modelo de regresión lineal solamente con intercepto.

- Para árboles de regresión de $m=1$ a M :

Se computan los residuos del modelo:

$$r_{mi} = y_i - f_{m-1}(x_i)$$

i corresponde a cada observación y

f_{m-1} corresponde a la suma de todos los árboles de regresión previos.

- Posteriormente se ajusta un árbol para los residuos.
- Para cada nodo terminal se computa la media residual.
- Se suma el árbol de regresión de los residuos al mejor ajuste corriente:

$$f_m = f_{m-1} + \text{último árbol de regresión de los residuos.}$$

A.4 Metodología K-means

El objetivo de k-means es dividir las observaciones en k-clusters. Por supuesto, es una clasificación cualitativa. Dado un conjunto de observaciones $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, donde cada una corresponde a un vector (dado que se toma en cuenta el valor observado para varias variables conjuntamente), el método de clusterización procura obtener $k \leq n$ grupos para las observaciones, siendo los grupos S_j , $j=1, 2, \dots, k$.

La meta del algoritmo es minimizar la suma de los cuadrados dentro del cluster (WCSS – Within Clusters Sum of Squares):

$$\operatorname{argmin}_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

donde $\boldsymbol{\mu}_i$ corresponde al centroide del grupo S_i y $\|\mathbf{x} - \boldsymbol{\mu}_i\|$ es la distancia euclidiana entre la observación y el centroide.

El primer paso consiste en seleccionar k centroides, uno para cada cluster. Si éstos no son especificados, el algoritmo los extrae aleatoriamente. En caso de querer seleccionarlos, conviene que se encuentren lo más alejados posibles entre ellos, por ejemplo, dividiendo los datos en percentiles y seleccionándolos en base a esa información.

Luego se van tomando de a una todas las observaciones, asociándolas al centroide más cercano.

Se recalculan los centroides, a partir de una media de todos los puntos que pertenecen al cluster:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

El algoritmo hace entonces un loop repitiendo los dos pasos anteriores hasta lograr la convergencia. Se detiene cuando la asignación no cambia de una iteración a la siguiente.

Tablas:

Tabla A.1
Descripción de las variables utilizadas

Nombre de la Variable	Descripción
Edad	Edad del individuo
Educación	Cantidad de años de educación
Género	Variable que toma valor 1 si es mujer y 0 si es hombre
Cantidad de Trabajos	Cantidad de trabajos que tiene el individuo
Aporta a Caja	Variable que toma valor 1 si aporta a alguna caja y 0 si no aporta
Cobra Aguinaldo	Variable que toma valor 1 si cobra aguinaldo y 0 si no cobra
Horas Trabajo Hogar	Cantidad de horas que dedica al trabajo en el hogar
Años Trabajo Actual	Cantidad de años que hace que trabaja en el empleo actual
Tamaño de la Empresa	Variable que toma valores de 1 a 6*
Trabaja en Local de la Empresa	Variable que toma valor 1 si trabaja en el local de la empresa y 0 si no
Desocupado Últimos 12 Meses	Variable que toma valor 1 si estuvo desocupado en los últimos 12 meses y 0 si no
Busca Trabajo	Variable que toma valor 1 si se encuentra buscando trabajo para sustituir y 0 si no
Medio de Transporte utilizado para ir al trabajo	Variable que toma valores de 1 a 7**
Situación Educativa	Variable con las 510 categorías de situación educativa creadas como se especifica en la metodología
Barrio	Variable que contiene los barrios de Montevideo urbano
CIIU	Variable que toma los valores de la Clasificación Internacional Industrial Uniforme
Ingreso	Ingreso de la ocupación principal menos aguinaldo y salario vacacional

Fuente: Elaboración Propia

Tabla A.2
Categorías de variable Tamaño Empresa (*)

Categoría	Detalle
1	1 empleado
2	2 a 4 empleados
3	5 a 9 empleados
4	10 a 19 empleados
5	20-49 empleados
6	50 o más empleados

Fuente: Elaboración Propia

Tabla A.3
Categorías de variable Medio de Transporte utilizado para ir al trabajo (**)

Categoría	Detalle
1	Transporte colectivo
2	Taxi o similar
3	Automóvil particular
4	Ciclomotor
5	Bicicleta
6	A pie
8	No se traslada
7	Otros

Fuente: Elaboración Propia

Tabla A.4
Estadísticas descriptivas de las variables utilizadas (N=378.410)

Variable	Mean	Std. Dev.	Min	Max
Edad	38	12	15	88
Educación	11	4	0	31
Género	0,47	0,49	0	1
Cantidad de Trabajos	1	0,42	1	9
Aporta a Caja	1	0,13	1	2
Cobra Aguinaldo	1	0,13	1	2
Horas Trabajo Hogar	0,70	3	0	60
Años Trabajo Actual	9	10	0	51
Tamaño de la Empresa	5	0,68	5	7
Trabaja en Local de la Empresa	1	0,24	0	2
Desocupado Últimos 12 Meses	2	0,27	1	2
Busca Trabajo	0,09	0,29	0	1
Medio de Transporte	2	2	1	8
Cluster Situación Educativa	5	2	1	7
Cluster Barrio	5	2	1	7
Cluster CIU	3	1	1	4
Cluster Situación Educativa 2	3	1	1	5
Cluster Barrio 2	3	1	1	6
Cluster CIU 2	4	0,49	1	4
Logaritmo de Ingreso	4,79	0,77	0,68	8,01