

Predicción de transiciones laborales de la actividad a la
inactividad en el mercado laboral argentino:
un enfoque de aprendizaje automático

Tesis de Maestría en Economía - Universidad Nacional de La Plata

Alumno: Agustin Staudt

Directora: María Inés Berniell

Codirector: Ramiro Heraclio Gálvez

2022

Resumen

El presente trabajo se propone comparar la capacidad predictiva que los modelos tradicionales y de machine learning tienen para pronosticar la probabilidad de salida del mercado de trabajo en Argentina, a fin de encontrar el estimador de mayor consistencia y poder predictivo. A su vez, se indaga acerca de la influencia que las características (individuales, familiares y laborales) de las personas activas tienen sobre dicha probabilidad, poniendo especial énfasis en entender el rol del género como determinante de las transiciones laborales. Para esto se utilizan datos de la Encuesta Permanente de Hogares para el período 2004-2018. Los resultados encontrados indican una diferencia estadísticamente significativa del modelo XGBoost a la hora de predecir salidas del mercado laboral respecto a los demás modelos, aunque también se destaca la competitividad de la regresión logit. En cuanto al género, se encuentra que su inclusión como variable es igual o más importante que aumentar la sofisticación del modelo. Ser mujer activa aumenta de manera importante las probabilidades de salir de la fuerza laboral en el siguiente período, en comparación a su contraparte masculina.

Palabras clave: Mercado laboral, Transiciones laborales, Salida de la fuerza laboral, Predicción, Género, Machine learning.

Abstract

This paper aims to compare the predictive capacity that traditional and machine learning models have to forecast the probability of going out-of-the labor force in Argentina, in order to find the most powerful and consistent estimator. In turn, the influence that active people's characteristics (as individual, family and labor) have on that probability are investigated, focusing on understanding gender as a labor transition determiner. To this end, data from the Encuesta Permanente de Hogares (EPH) for the 2004-2018 period are used. The results show a statistically significant difference in predicting labor market exits between XGBoost and the other models (including in this paper), although the competitiveness of the logit regression is also highlighted. Regarding gender, the results point out that its inclusion as a variable is at least as important as increasing the sophistication of the model. Being an active woman increases considerably the probabilities of leaving the labor force (market) in the next period, compared to her male counterpart.

Keywords: Labor market, Labor transition, Out-of-the labor force, Prediction, Gender, Machine learning.

1 Introducción

El mercado laboral se encuentra en continuo movimiento interno, con trabajadores transitando rápidamente entre diferentes trabajos, hacia y desde el desempleo, o sin participar incluso en momentos donde la actividad económica se encuentra estable (Tasci y Tansel, 2005). En el curso de su vida laboral activa los individuos entran y salen del mercado de trabajo, sin embargo, dichas transiciones no son iguales para todos los grupos poblacionales. La movilidad entre trabajos, la posibilidad de mantenerse en el mercado laboral o salir sistemática y permanentemente de él no es indiferente al hablarse de mujeres y varones, de jóvenes, o de personas con mayor y menor educación (Theodossiou, 2002; Theodossiou y Zangelidis, 2009; Nilsen et al., 2000; Cerrutti, 2000; Marston et al., 1976). Esta heterogeneidad conlleva diferentes oportunidades de obtener empleos regulares, distintos efectos sobre el nivel de experiencia general y específica que se acumula dentro de la fuerza laboral, diferencias en los niveles de inversión en capacitación laboral, tanto por parte de la firma como por parte del trabajador, como también, conlleva diferentes niveles subsecuentes en el ingreso, imposibilitando en muchos casos que las personas puedan permanecer y desarrollarse plenamente en el mundo del trabajo (Keith y McWilliams, 1997; Theodossiou y Zangelidis, 2009; Marchionni et al., 2019).

La literatura que investiga patrones de movilidad laboral es amplia. En países desarrollados Marston et al. (1976) evalúa los flujos entre diferentes estados laborales para Estados Unidos; Nilsen et al. (2000) estiman transiciones desde el empleo para Noruega; Theeuwes et al. (1990) estudian diferentes transiciones laborales para Holanda; Theodossiou (2002) lo hace para el Reino Unido; y Theodossiou y Zangelidis (2009) para seis países pertenecientes a la comunidad europea (Reino Unido, Francia, Alemania, Finlandia, España y Grecia). En el caso de países en desarrollo, Tasci y Tansel (2005) realiza un análisis sistemático de los determinantes de movilidad para el mercado laboral de Turquía; Woltermann (2002) lo realiza para Brasil; CEDLAS y BID (2019) se enfoca en el análisis de intermitencia laboral para mujeres pertenecientes a México y Perú; y Cerrutti (2000) analiza la movilidad para mujeres del mercado laboral argentino.

A partir del amplio abanico de trabajos que estudian los patrones de movilidad laboral surge la pregunta de si existiría una manera de predecir los cambios futuros de estado ocupacional dentro y fuera del mercado de trabajo. Esto permitiría anticiparse a probables transiciones dentro de diferentes grupos poblacionales, identificando a los candidatos más

factibles de realizar movimientos hacia y desde la fuerza laboral. Dicho análisis contribuiría tanto a la literatura de movilidad como a la generación de políticas públicas. En cuanto a la contribución a la literatura, la predicción de transiciones permitiría caracterizar y entender la dinámica existente de cambios en los estados laborales de personas que participan en el mercado. Por el lado de la política pública, posibilitaría el diseño e implementación de medidas focalizadas que puedan atacar cada problemática puntual de la manera más efectiva, minimizando el riesgo de generación de transiciones involuntarias o forzadas que impidan la participación y desarrollo pleno de cada trabajador.

En Argentina, la Encuesta Permanente de Hogares (EPH) es el mayor relevamiento de información sociolaboral de la actualidad en el país, la cual permite llevar a cabo este tipo de análisis predictivo de transiciones laborales. Por lo que el presente trabajo se propone estudiar la viabilidad de desarrollar modelos que predigan de la manera más certera posible la transición laboral de corto plazo en el mercado laboral argentino usando la EPH. Puntualmente, este estudio se enfoca en el análisis de salida del mercado laboral, es decir, en predecir la probabilidad de transición de la actividad a la inactividad de un año a otro. Para ello, se busca el modelo de mayor consistencia y poder predictivo, y se indaga acerca de la influencia que las características (tanto individuales como familiares y laborales) de las personas activas tienen sobre dicha probabilidad. Para llevar a cabo el análisis predictivo se utilizan dos métodos. En primer lugar, se realiza la estimación a partir del modelo de regresión logística (o logit), el cual es considerado en el trabajo como un enfoque tradicional dentro de la literatura de transiciones laborales, debido a su frecuente implementación.¹

En segundo lugar, se agregan al análisis dos modelos de clasificación de aprendizaje automático (ML), el estimador lasso y el método XGBoost (*Extreme Gradient Boosting*). Debido a las ventajas que poseen los modelos de aprendizaje supervisado a la hora de predecir, se considera fundamental agregar al análisis estimadores más modernos, con el objetivo de comparar la *performance* predictiva entre un método comúnmente utilizado por trabajos previos de movilidad laboral (regresión logit), con uno que ha tomado mayor relevancia en los últimos años por distintos campos de la economía, y aún poco utilizado por la literatura de transiciones laborales. Los modelos ML plantean alternativas desarrolladas específicamente para maximizar el rendimiento de predicción, ya sea desde proponer

¹ La utilización de la regresión logit como metodología de análisis puede encontrarse en trabajos como los de [Tasci y Tansel \(2005\)](#), [CEDLAS y BID \(2019\)](#), [Nilsen et al. \(2000\)](#), entre otros.

estructuras o formas más flexibles que los métodos tradicionales, hasta cambiar la manera de utilizar los datos para mejorar la precisión en la predicción de nuevas observaciones (Varian, 2014; James et al., 2013; Mullainathan y Spiess, 2017; Kleinberg et al., 2015).

En este estudio se pone especial énfasis en entender el rol del género como determinante de las transiciones laborales. La evidencia previa documentada por la literatura de transiciones encuentra significativas diferencias de género, las mujeres parecen tener más chances de abandonar el mercado laboral, en relación a los varones. Las restricciones sociales que asocian a las mujeres con el trabajo doméstico y de cuidado en muchas ocasiones impiden que puedan gozar de una participación activa y prolongada en el mundo del trabajo, en comparación a su contraparte masculina (Hersch y Stratton, 1997; Theodossiou, 2002; Theodossiou y Zangelidis, 2009). A diferencia de otros factores comúnmente encontrados en trabajos previos (como educación, experiencia e ingresos), la presencia de brechas de género en transiciones hacia y desde la fuerza laboral es una situación no deseable en pos de lograr un desarrollo sostenible a largo plazo.² Por lo que se considera importante enfocar el análisis en las diferencias que existen en transiciones de varones y mujeres, como también, cuán relevantes son las mismas en relación a otros determinantes que podrían influir en la salida del mercado laboral.

Los resultados encontrados indican una diferencia estadísticamente significativa del modelo XGBoost a la hora de predecir salidas del mercado laboral respecto a los demás modelos.³ Aunque se remarca la *performance* predictiva del modelo logit, ya que la diferencia en el desempeño entre este y el modelo XGBoost es baja, coincidiendo así con los hallazgos documentados previamente (Salganik et al., 2020; Bhattacharyya et al., 2011; Whitrow et al., 2009; Pesantez-Narvaez et al., 2019; Khanna et al., 2015; Jang Su-Kyeong, 2020; Lee et al., 2018). Al realizar las predicciones para una especificación sin la variable de género, el modelo logístico presenta un AUC-ROC de 0.81 en el grupo de validación,⁴ en

² Si bien no se desarrolla en el trabajo, existe una amplia literatura que aborda la cuestión de desigualdad de oportunidades (IO). Argumentando que no toda desigualdad es mala, sino aquella atribuible a circunstancias predeterminadas, es decir, que no se pueden elegir por venir preestablecidas (género, etnia, lugar de nacimiento, características de los padres, etc.). Así, siguiendo esta línea de pensamiento, las diferencias entre varones y mujeres son una condición injusta, no así factores que podrían deberse al esfuerzo como el ingreso o la cantidad de horas trabajadas. Por supuesto, los resultados desarrollados en el trabajo se deben tomar con cautela, debido a la posible correlación entre cada uno de los factores bajo estudio, que impide aislar el efecto del género respecto al de otras características de los activos. Para más información de la literatura de IO puede consultarse a Ferreira y Gignoux (2014).

³ La significatividad estadística para comparar el desempeño predictivo de cada modelo se computa a partir del *p-value* propuesto en DeLong et al. (1988).

⁴ A la hora de realizar la predicción es común implementar el enfoque de validación, lo cual significa separar los datos en tres subconjuntos distintos. Un set de entrenamiento o training para estimar el

lasso este valor es 0.80, mientras que el modelo XGBoost la *performance* es de 0.835, es decir, el modelo XGBoost tiene casi un 84% de probabilidades de que si se eligen dos observaciones al azar, una de “Salida laboral” y otra de “No Salida”, la probabilidad predicha para la “Salida” sea mayor que para la “No salida”. Un valor AUC equivalente a casi 3 puntos porcentuales más que logit.⁵ La inclusión de la variable de género aumenta significativamente la *performance* predictiva de los tres modelos (0.834 en logit, 0.827 lasso y 0.85 en xgboost), remarcando así la importancia de considerar a esta para predecir salidas del mercado laboral. Por otra parte, existe un error de predicción no despreciable que se da en los tres modelos, resultado que se alinea con los hallazgos de [Salganik et al. \(2020\)](#), quienes explican que dicho error estaría fuertemente relacionado con la dificultad de predecir patrones sociales y débilmente asociado con la técnica usada para generar la predicción. Por último, el mejor modelo proyecta una salida del 13% de individuos activos en el 2017, a generarse en el siguiente año, de los cuales el 87% son mujeres activas.

De lo anterior se destacan dos puntos importantes. Por un lado, el cambio de un modelo tradicional (logit) a uno más sofisticado (XGBoost) para predecir salidas del mercado laboral aumenta significativamente la *performance* predictiva en 2 puntos porcentuales. Por otro lado, los resultados sugieren que la inclusión del género como variable es igual o más importante que aumentar la sofisticación del modelo. Respecto a esto último, a través de distintos métodos de selección de variables la *dummy* de género se encuentra entre las variables más influyentes de la predicción. En el caso de logit es la segunda más importante, mientras que en la regresión lasso es seleccionada en primer lugar. En el caso de XGBoost, de los dos métodos utilizados uno lo ubica en segundo lugar (método de Ganancia) y el otro en el primer lugar (importancia SHAP). En la mayoría de los modelos, el género está por delante o detrás del ingreso de la ocupación principal para los trabajadores, es decir, que considerar el género es tan importante como considerar el ingreso laboral de una persona.

Por supuesto, como anticipa la literatura, la presencia de brechas de género en la transición

modelo, un subconjunto de validación para elegir el modelo y un subconjunto test para evaluar cuán bien se desempeña el modelo elegido ([Varian, 2014](#)).

⁵ La curva *Receiver Operating Characteristic*(ROC) es un gráfico que muestra el rendimiento de un modelo de variable de elección binaria para todos los umbrales de clasificación (o puntos de corte). Así el Área bajo la curva ROC (o AUC-ROC) mide toda el área bidimensional por debajo de la curva ROC y se interpreta como la probabilidad de que un clasificador ordene una instancia positiva elegida aleatoriamente más alta que una negativa. Dicha métrica representa una medida típicamente utilizada por la literatura de ML, y cuya explicación detallada será dada más adelante. Cabe aclarar que la definición presentada se obtuvo de [Google](#).

va en detrimento de la mujer. Las mujeres en condición de actividad laboral tienen mayor probabilidad de salir del mercado de trabajo en el siguiente año en relación a los varones. El hecho de ser mujer aumenta casi 7 puntos porcentuales la probabilidad de salida respecto a su contraparte masculina. A su vez, al interactuar el género con variables como la situación conyugal o la presencia de niños menores de 10 años, la asociación es positiva, es decir, ser mujer casada o vivir en un hogar con al menos un niño menor de edad aumenta la probabilidad de salida.

Estos resultados son compatibles con lo encontrado por trabajos previos, que remarcan la existencia de roles heterogéneos dentro del hogar, donde la mujer dedica la mayor parte de su tiempo en tareas domésticas y de cuidado bajo un rol secundario en la generación de ingresos para su familia, característica que se asocia con una mayor predisposición a salir del mundo del trabajo cuando las obligaciones familiares lo requieran. Mientras el varón participa en el mercado laboral como el principal generador de ingresos del hogar (Keith y McWilliams, 1997; Hersch y Stratton, 1997; Cerrutti, 2000; Theodossiou, 2002; Theodossiou y Zangelidis, 2009; Marchionni et al., 2019). Adicionalmente, analizando de forma descriptiva las transiciones opuestas, es decir, inserción al mercado de trabajo desde la inactividad, se encuentra que la salida a corto plazo del mercado laboral de las mujeres no se condice, en su mayoría, con un futuro retorno inmediato a la actividad. Por lo que las salidas podrían ser incluso permanentes.

Además del género, se encuentran otros factores relevantes que se alinean con los hallazgos de la literatura de movilidad laboral. La edad, educación, formalidad, antigüedad en el puesto e ingresos se relacionan negativamente con la probabilidad de salir de la fuerza de trabajo. Siguiendo el análisis de trabajos previos, la posibilidad de acumular experiencia, habilidades, educación e ingresos permite tener mayor estabilidad y reducir así la chance de interrumpir los vínculos laborales (Theeuwes et al., 1990; Nilsen et al., 2000; Tasci y Tansel, 2005).

A través del análisis predictivo el trabajo contribuye a la literatura de transiciones laborales en tres direcciones. En primer lugar, se provee el modelo con mayor poder predictivo para estimar transiciones de actividad a la inactividad. En segundo lugar, se analiza la importancia relativa del género para predecir dichas salidas del mercado de trabajo. Por último, a partir de la búsqueda del mejor estimador se comparan dos enfoques de predicción, el método tradicional con el de machine learning. Los hallazgos de este trabajo

serían de utilidad no sólo para la literatura de transiciones laborales, sino también para la política pública, debido a que los resultados permitirían anticipar el flujo de salida del mercado laboral que podría darse en el siguiente período, como también, se podrían detectar los candidatos más factibles de pasar a la inactividad, lo cual ayudaría a mejorar el diseño e implementación de políticas públicas focalizadas.

En lo restante del trabajo, se inicia el análisis con la discusión de los principales resultados de la literatura de transiciones laborales de la actividad a la inactividad, como también, se estudian los trabajos que comparan ambos enfoques de predicción bajo estudio (Sección 2). En la sección 3 se describe la fuente de datos, metodología y se analizan exploratoriamente los datos de entrenamiento. En la sección 4 se muestran los resultados de los diversos modelos de predicción implementados. Finalmente, en la sección 5 se presentan las conclusiones del trabajo.

2 Revisión de la literatura

2.1 Transiciones laborales a la inactividad

Estudios previos han documentado diferentes determinantes relacionados con la probabilidad de salir de la fuerza laboral. Entre los hallazgos más comunes de la literatura se encuentran variables vinculadas con el mercado laboral, como experiencia, y nivel de ingresos; características personales como la situación conyugal, presencia de menores en el hogar, educación y género; y variables regionales como la tasa de desempleo local (Nilsen et al., 2000; Theeuwes et al., 1990; Theodossiou y Zangelidis, 2009; Theodossiou, 2002).

En el caso de las características laborales individuales la literatura encuentra que tanto los años de experiencia, la edad, nivel educativo y el nivel de ingresos se asocian positivamente con la probabilidad de mantenerse en el mercado laboral (Theodossiou y Zangelidis, 2009; Theodossiou, 2002; Theeuwes et al., 1990; Tasci y Tansel, 2005; Nilsen et al., 2000; Marston et al., 1976). Por su parte, condiciones regionales o locales afectan la situación laboral de los activos. Si bien la condición local no es un factor analizado en profundidad, diferentes autores encuentran una relación positiva entre la tasa de desempleo local y la probabilidad de salir del mercado laboral, es decir, aquellos trabajadores que viven en zonas con alto nivel de desempleo tienen mayor probabilidad de salir del mercado de trabajo una vez que pierden su empleo o se encuentran buscando uno (Theeuwes et al., 1990; Tasci y Tansel,

2005).

Muchos de los determinantes encontrados por la literatura de transiciones laborales remarcan importantes diferencias de género. Según [Theodossiou \(2002\)](#), las mujeres tienen en promedio un riesgo mayor de salida o despido, muestran un menor compromiso con la actividad del mercado laboral y son relativamente menos móviles que los varones. También, es más probable que abandonen la fuerza de trabajo por motivos familiares o personales ([Keith y McWilliams, 1997](#); [Marston et al., 1976](#)).

Estudios previos también encuentran efectos heterogéneos de género según el tipo de ocupación y la categoría ocupacional. A su vez, las características como la duración en el desempleo o el método de búsqueda laboral no afecta de la misma forma a mujeres y varones ([Tasci y Tansel, 2005](#)). No obstante, las diferencias más notorias de género documentadas por la literatura de transiciones laborales se encuentran en variables que resumen la situación dentro del hogar. Factores como la situación conyugal y la presencia de niños menores afecta considerablemente la probabilidad de las mujeres de salir de la fuerza laboral, algo que no ocurre en los varones y de ocurrir la magnitud del efecto es significativamente menor ([Theeuwes et al., 1990](#); [Tasci y Tansel, 2005](#); [Theodossiou y Zangelidis, 2009](#); [Hersch y Stratton, 1997](#); [Marston et al., 1976](#); [Tasci y Tansel, 2005](#); [CEDLAS y BID, 2019](#)).

Si bien la literatura de transiciones laborales de la actividad a la inactividad laboral es amplia, muchos de los trabajos citados analizan determinantes que se observan en el momento t_0 y su relación con la probabilidad de salida en el mismo momento observado t_0 . El presente trabajo se distingue en que se observan las características del individuo activo en el momento t_0 , para predecir la probabilidad de salida del mismo en el momento futuro t_1 . Este último enfoque resulta aún poco abordado a pesar de su relevancia a la hora de elaborar políticas que puedan anticiparse a flujos futuros de salida del mercado laboral.

Por otra parte, las considerables diferencias sistemáticas existentes entre mujeres y varones a la hora de estudiar las transiciones hacia la inactividad, remarcan la necesidad de considerar como factor central las brechas de género para el estudio de transiciones laborales, debido a que a diferencia de factores como el ingreso o educación, la presencia de desigualdades entre mujeres y varones es una condición injusta, en el sentido de que el género es una situación predeterminada, y por lo tanto no deseable en pos de lograr un desar-

rollo sostenible a largo plazo. Por lo tanto, en el análisis del presente trabajo se pretende dedicar gran parte del estudio a evaluar y describir la importancia relativa que tiene el género para predecir las probabilidades de salida del mercado laboral bajo el ejercicio predictivo propuesto.

2.2 Comparación modelos de machine learning versus modelos tradicionales

Cuán bien se desempeñan los modelos ML y las técnicas de regresión a la hora de predecir? Diferentes trabajos evaluaron dicha cuestión, aunque la mayoría pertenecientes a campos ajenos a cualquier área dentro de la economía laboral. Quizás el más vinculado a cuestiones del mercado de trabajo es el análisis de [Salganik et al. \(2020\)](#), quienes investigan la viabilidad de predecir trayectorias de vida, incluyendo probabilidad de despido laboral. Los hallazgos de este trabajo indican que, en promedio el cambio de un modelo tradicional simple a uno de mayor poder predictivo disminuye el error cuadrático medio (MSE) en 0.04 puntos para el grupo test. Concluyendo así que si bien los modelos de machine learning aumentan el poder predictivo respecto a los métodos tradicionales, estos últimos son considerablemente competitivos. A su vez, los autores encuentran errores de predicción no menores, independientemente de la técnica a utilizar. Por lo que [Salganik et al. \(2020\)](#) explican que este error está fuertemente relacionado con la dificultad de predecir patrones sociales y débilmente asociado con la técnica usada para generar la predicción.

Los demás trabajos que comparan ambos enfoques se encuentran en diferentes áreas poco o nada relacionadas a cuestiones del mercado laboral y muy diferentes entre sí. Trabajos dentro del sector financiero ([Bhattacharyya et al., 2011](#); [Whitrow et al., 2009](#)); cuestiones viales como prevenir accidentes de tránsito ([Pesantez-Narvaez et al., 2019](#)); como también, trabajos dentro del sector de salud ([Khanna et al., 2015](#); [Jang Su-Kyeong, 2020](#); [Lee et al., 2018](#)). En general, los resultados encontrados remarcan hallazgos similares a [Salganik et al. \(2020\)](#), es decir, los métodos de ML mejoran significativamente la predicción aunque no difieren considerablemente de los modelos tradicionales, remarcando la competitividad de estos últimos. En definitiva, la comparación del desempeño predictivo de ambos enfoques es escasa a la hora de predecir transiciones laborales. Por lo que, el presente trabajo se propone colaborar con la literatura realizando un análisis comparativo para predecir salidas futuras del mercado laboral.

3 Fuente de datos y metodología

3.1 Fuente de datos

Para realizar el trabajo se utilizó como insumo la Encuesta Permanente de Hogares (EPH), el mayor relevamiento de información sociolaboral de la actualidad en Argentina, el cual es representativo de la población urbana de los 31 aglomerados más grandes del país. El período de análisis se encuentra entre el primer trimestre de 2004 y el cuarto trimestre de 2019. A su vez, las variables monetarias se encuentran deflactadas por el Índice de Precios al Consumidor (IPC) y expresadas en pesos del cuarto trimestre de 2019.

La EPH posee un seguimiento bajo la modalidad 2-2-2, es decir, a partir del primer trimestre donde el hogar ingresa al relevamiento (llámese período t_0) este vuelve a ser encuestado en el trimestre contiguo, sale temporalmente por los siguientes dos, para volver a ingresar a la encuesta en el mismo trimestre que fue encuestado por primera vez pero del siguiente año, luego se repite su relevamiento por cuarta y última vez el trimestre siguiente (llámese a este último período t_1). Totalizando así un seguimiento de año y medio.⁶

A partir de la información relevada de cada individuo en el momento t_0 se construyen los predictores individuales y del hogar (los cuales se detallan en el Anexo A). En cuanto al *outcome* de interés, es decir, la *dummy* que indica si la persona activa pasa o no a la inactividad en el siguiente período utilizamos el estatus laboral de la persona en t_0 y en t_1 . Cabe aclarar que se debe considerar la posibilidad de que los hogares salgan de la encuesta antes de cumplir el período completo de seguimiento, por lo que se establece como condición adicional de que, entre el período inicial y final hayan pasado al menos 100 días, de lo contrario el hogar no sería considerado dentro de las observaciones. De esta forma, el presente trabajo estudia la probabilidad de transición de corto plazo, esto es, la probabilidad de movilidad laboral de actividad a inactividad en un período no mayor al año y medio. Por último, para evitar considerar salidas del mercado laboral por razones de estudio o asociadas a la edad legal de retiro, se excluye de la muestra a los menores de 25 años y a los mayores de 59 años. Así, la población bajo estudio es considerada en el presente trabajo como las personas en edad efectiva de trabajo.

⁶ Para ilustrar el procedimiento del seguimiento pensemos en el ejemplo de un hogar que ingresa a la encuesta en el primer trimestre de 2014 (t_0). Este vuelve a ser encuestado en el segundo trimestre de ese mismo año, sale temporalmente el tercer y cuarto trimestre de 2014, para luego volver a ingresar el primero de 2015. Por último, este hogar es encuestado por cuarta y última vez en el segundo trimestre de 2015 (momento t_1).

Cabe remarcar que existen dos inconvenientes relacionados al seguimiento de la misma persona. En primer lugar, puede pasar que el individuo identificado en el momento t_0 no sea la misma que en el momento t_1 , por lo que se agregan condiciones adicionales para que el encuestado forme parte de la base, esto es, tener el mismo sexo entre ambos períodos y la diferencia de edad entre el momento t_0 y t_1 debe ser menor a 2 años.⁷ En segundo lugar, puede ocurrir que, aunque el hogar continúe en la encuesta en ambos momentos, algún integrante no se encuentre en alguno de los dos períodos, por lo que solamente se consideran a aquellos individuos que estén tanto en el período t_0 como en el período t_1 , independientemente de la situación del hogar. De esta forma, al contemplar los filtros explicados antes y eliminando las observaciones con valores *missing* el *dataset* se conforma de 191378 personas activas de 25 a 59 años de edad correspondientes al período 2004-2018 y 56 variables predictoras detalladas en el Anexo A.^{8 9}

Adicionalmente, se incluyen distintas series agregadas con la intención de captar los ciclos económicos de cada trimestre en Argentina. Para ello se suma a la base de datos el Estimador Mensual de la Actividad Económica (EMAE a partir de ahora), el cual es realizado por el Instituto Nacional de Estadísticas y Censo de Argentina (INDEC), el Tipo de Cambio Nominal Promedio Mensual (TCNPM) elaborado por el Banco Central de la República Argentina y el Índice Líder de la Universidad Torcuato Di Tella.¹⁰

3.2 Metodología

Siguiendo el análisis de [Pesantez-Narvaez et al. \(2019\)](#) para un conjunto de datos de N observaciones y P covariables, tenemos una variable de elección binaria y_i con $i = 1, \dots, n$ que toma valores 0,1; y un conjunto de covariables denotadas como X_{ip} con $p = 1, \dots, P$. La función de densidad de probabilidad condicional de $y_i = t (t = 0, 1)$ dado $X_i (X_{i1}, \dots, X_{iP})$, es denotado como $\pi_t(X_i)$. Equivalentemente, esto último puede ser $Prob(y_i = t) = \pi_t(X_i)$, tal que $E(y_i) = Prob(y_i = 1) = \pi_1(X_i)$.

⁷ Para el caso de los recién nacidos (personas con edad -1) su inclusión a la base completa (previo al filtro por edad) se realiza si su diferencia de años entre ambos períodos no es mayor a 3. Dicha subpoblación representa el 10% del total de individuos entre 25 y 59 años de edad.

⁸ Si bien la EPH continúa se inicia el tercer trimestre de 2003, la inclusión de variables macroeconómicas genera que el período de análisis comience en 2004. Por otra parte, las características observables llegan hasta 2018, sin embargo, para armar la condición laboral futura de ese año se utiliza información del estado ocupacional de 2019.

⁹ Tanto el procesamiento de la base de datos como el análisis realizado en todo el trabajo se realiza utilizando el software *open-source* R y Rstudio de [R Core Team \(2021\)](#).

¹⁰ Para más información del EMAE puede consultar [INDEC](#), en cuanto al tipo de cambio se puede consultar [BCRA](#), mientras que información sobre el Índice Líder se encuentra disponible en [UTDT](#).

3.2.1 Regresión Logística

El modelo de regresión logística ha sido uno de los métodos más utilizados por la literatura de transiciones laborales.¹¹ Dicho estimador utiliza la función logit como un enlace canónico, es decir, el logaritmo del ratio de las funciones de probabilidad $\pi_i(X_i)$ es una función lineal de X , esto es:

$$\log \frac{\pi_1(X_i)}{\pi_0(X_i)} = \log \frac{Prob(y_i = 1)}{Prob(y_i = 0)} = \beta_0 + \sum_{p=1}^p X_{ip}\beta_p, \quad (1)$$

donde $\beta_0, \beta_1, \dots, \beta_p$ son los parámetros del modelo, $Prob(y_i = 1)$ es la probabilidad de observar el evento en la variable binaria ($y_i = 1$), y $Prob(y_i = 0)$ es la probabilidad de no observarlo ($y_i = 0$). A efectos de interpretación, la Ecuación 1 puede ser escrito en términos de las probabilidades de suceso del evento ¹²

$$\frac{\pi}{1 - \pi} = \exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p) \quad (2)$$

Dada la ecuación 2 podemos establecer lo siguiente

$$\pi = \frac{\exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p)}{1 + \exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p)} \quad (3)$$

Suponiendo que los x_p no están funcionalmente correlacionados con $p = 1, 2, \dots, p$ el efecto del cambio en una unidad en x_1 es el cambio del *log odds* por una cantidad β_1 . Equivalentemente, se podría decir que el efecto de un cambio de una unidad en x_1 es el incremento de la probabilidad de suceso del evento multiplicado por el factor $\exp(\beta_1)$.¹³

Los modelos logit se optimizan regularmente a través de la función de máxima verosimilitud (*log-likelihood*).¹⁴

¹¹ Para la formalización del modelo logit se siguen los trabajos de [Pesantez-Narvaez et al. \(2019\)](#), [McCullagh y Nelder \(1989\)](#) y [Hastie et al. \(2009\)](#).

¹² Para simplificar la notación se reescribe $\pi_1(X_i) = \pi$ y $\pi_0(X_i) = 1 - \pi$.

¹³ Como explican [McCullagh y Nelder \(1989\)](#) es importante que los x_k con $k \neq 1$ se mantengan fijos y no se permitan variar como consecuencia de un cambio en x_1 .

¹⁴ La derivación de la Ecuación 4 se incluye en el Anexo B.1.

$$\ell(\theta) = \sum_{i=1}^N \left\{ y_i(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p) - \log(1 + \exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p)) \right\} \quad (4)$$

3.2.2 Regresión lasso

El modelo tradicional tiene diferentes ventajas en términos de inferencia y, en los problemas del mundo real, es frecuentemente competitivo en relación a diferentes métodos no lineales (James et al., 2013). Por lo que antes de pasar al mundo de la no linealidad uno se podría preguntar si a partir de algún procedimiento alternativo, este método podría modificarse de forma tal de mejorar el poder predictivo y la interpretabilidad.¹⁵ En el caso del poder predictivo, cuando la cantidad de observaciones N no es considerablemente mayor que p , podríamos encontrarnos en un escenario de mucha variabilidad en la estimación logit, resultando así en un sobreajuste (*overfitting*) y consecuentemente en predicciones pobres para observaciones nuevas, no utilizadas para entrenar el modelo. Por lo que restringiendo o contrayendo los coeficientes estimados, se podría reducir sustancialmente la varianza a costa de un incremento leve en el sesgo del estimador. Lo cual podría generar mejoras en el poder predictivo del modelo a la hora de predecir datos desconocidos (James et al., 2013).

En cuanto a la interpretabilidad, la irrelevancia de predictores dentro del modelo logit lleva a una complejidad innecesaria en el estimador resultante. Por lo que, utilizando alguna técnica de selección de variables se podría obtener un modelo más fácil de interpretar en términos predictivos. Dejando aquellos predictores que influyen en la estimación, y sacando los que no la “empujan”.

Por lo que el trabajo intenta mejorar los resultados de la regresión logística agregando un término de penalidad L_1 (lasso), la cual puede ser usada para seleccionar variables y contraer los coeficientes estimados. Así, se maximiza la siguiente versión penalizada de la Ecuación 4

$$\max_{\theta} \left\{ \sum_{i=1}^N \left[y_i(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p) - \log(1 + \exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p)) \right] - \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (5)$$

donde $\lambda \geq 0$ es un hiperparámetro de *tuning* que debe ser determinado separadamente de

¹⁵ El desarrollo de este apartado se hizo a partir de James et al. (2013) y Hastie et al. (2009).

la maximización.¹⁶ Cuando $\lambda = 0$, el término de penalidad no tiene efecto, por lo que la Ecuación 5 producirá estimaciones similares a logit. Sin embargo, a medida que $\lambda \rightarrow \infty$ el término $\lambda \sum_{j=1}^p |\beta_j|$ tiene el efecto de forzar los coeficientes estimados hacia exactamente cero. Por lo que ante un hiperparámetro lo suficientemente grande, la estimación logit penalizada produce una selección de variables.¹⁷

Cabe aclarar que, los coeficientes estimados a partir de la Ecuación 5 pueden cambiar sustancialmente al multiplicar un predictor por una constante (no son invariantes a la escala), por lo que estos son estandarizados a través de la siguiente fórmula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x})^2}} \quad (6)$$

de manera de que todos tengan la misma escala. El denominador de la Ecuación 6 es la desviación estándar estimada del j -ésimo predictor. Consecuentemente, todas las variables independientes estandarizados tendrán un desvío estándar igual a uno. Como resultado, el ajuste final no depende de la escala sobre la cual los predictores se midan (James et al., 2013).

3.2.3 Extreme Gradient Boosting (XGBoost)

La técnica de *Boosting* es una de las ideas de aprendizaje más eficaces de los últimos 20 años, cuya motivación ha sido la búsqueda de un procedimiento que combine *outputs* de muchos clasificadores “débiles” para producir un potente estimador agregado (Hastie et al., 2009). Dentro de este enfoque predictivo, Extreme Gradient Boosting (XGBoost) es reconocido como un algoritmo de amplio uso por su capacidad predictiva excepcional (Pesantez-Narvaez et al., 2019; Chen y Guestrin, 2016). La idea principal de este algoritmo es que el mismo construye D árboles de clasificación (CART) uno por uno, tal que el modelo subsecuente es entrenado utilizando el residual del árbol previo (*decision tree ensembles*).¹⁸

¹⁶ Al hablar de hiperparámetros el trabajo se refiere a aquellos componentes del modelo cuyos valores se establecen separadamente de la optimización, es decir, la persona los elige previamente. En cambio, los parámetros se obtienen durante el proceso de optimización, por lo que se establecen a partir del conjunto de datos.

¹⁷ Cabe aclarar que la penalidad se aplica a β_1, \dots, β_p , no al intercepto β_0 . Se pretende contraer la asociación estimada de cada variable con el *outcome*, sin embargo, no se quiere contraer el intercepto que es simplemente una medida de la media de la dependiente cuando $x_{i1}, \dots, x_{ip} = 0$ (James et al., 2013).

¹⁸ Un modelo CART es diferente de los árboles de decisión, debido a que los valores finales de cada nodo contiene valores de decisión. En CART, se asocia cada *score* con cada nodo final, lo cual genera una interpretación más rica. Para más información puede consultar XGBoost.

Esto es, el nuevo árbol corrige los errores hechos por el árbol previamente entrenado y predice nuevamente el *outcome*. Así, en XGBoost cada modelo ensamblado utiliza la suma de las D funciones para predecir el output

$$\hat{Y}_i = \sum_{d=1}^D f_d(X_i), f_d \in F, i = 1, \dots, N \quad (7)$$

Donde F es el conjunto de todos los modelos CART posibles, y cada f_d representa una estructura de árbol independiente que es denotada como q .¹⁹ Esto es, q representa el conjunto de reglas de un modelo que clasifica cada individuo i en un nodo final. La fase de entrenamiento implica clasificar N observaciones tal que, dado las variables X , cada nodo final tenga un *score* que corresponda a la proporción de casos que son clasificados dentro de la ocurrencia del evento para esa combinación específica de X_i . Se denota al *score* como $w_{q(X)}$.

Luego, q puede escribirse como una función $q : \mathbb{R}^P \rightarrow T$, donde T es el número total de nodos finales de un árbol y j denota un nodo final individual, $j = 1, \dots, T$. Para calcular la predicción final para cada individuo, el *score* de los nodos finales se suman mediante 7, donde $F = \{f(x) = w_{q(X)}\}$ con $q : \mathbb{R}^P \rightarrow T$ y $w \in \mathbb{R}^T$.

En general los métodos de *boosting* ajustan cada árbol a través de versiones re-ponderadas. La ponderación es un mecanismo que penaliza las predicciones incorrectas de los modelos anteriores, con el objetivo de mejorar la predicción. Así, la estructura de ponderadores son generalmente valores óptimos, que son ajustados una vez que la función de pérdida es minimizada. Por lo que los nuevos modelos incorporan la nueva estructura de ponderación en cada iteración, y predice un nuevo *outcome*.

La función a ser optimizada está dada por

$$\mathcal{L} = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{d=1}^D \Omega(f_d) \quad (8)$$

donde l es una función de pérdida convexa diferenciable que mide la diferencia entre la predicción \hat{y}_i y la dependiente y_i . Mientras que el segundo término es el componente que penaliza la complejidad de las funciones creadas por XGBoost.²⁰ Intuitivamente, la

¹⁹ Para el desarrollo de este apartado se utilizaron los trabajos de [Pesantez-Narvaez et al. \(2019\)](#) y [XG-Boost](#).

²⁰ La explicación del segundo término puede encontrarse en la Sección [B.2](#) del Anexo [B](#).

función objetivo regularizada tenderá a seleccionar aquel modelo que provea funciones predictivas y simples.

Finalmente, además de la penalización de complejidad, en *Extreme Gradient Boosting* existen una amplia variedad de otros hiperparámetros que se deben elegir previo a la ejecución del algoritmo.²¹ A continuación se listan los más comunes a establecer

- **nrounds**: número de árboles. Rango $[0, \infty)$.
- **max_depth**: profundidad máxima de los árboles. Rango $[0, \infty]$.
- **eta**: proporción a la que el árbol aprende. Rango $[0, 1]$.
- **gamma**: mínima reducción del error para generar un corte. Rango $[0, \infty]$.
- **colsample_bytree**: variables a muestrear y considerar en cada árbol. Rango $[0, 1]$.
- **min_child_weight**: mínima cantidad de observaciones en los nodos finales para considerar un corte. Rango $[0, \infty]$.
- **subsample**: muestreo de observaciones a considerar en cada árbol. Rango $[0, 1]$.

A pesar de la gran capacidad predictiva atribuida a XGBoost, el modelo dista de poseer interpretabilidad de los resultados, es decir, resulta difícil entender la forma en que el método utiliza cada una de los predictores disponibles.²² Es por eso que en los últimos años se ha propuesto una técnica para poder interpretar modelos complejos, conocido como *Shapley Additive Explanations* (SHAP).²³ Esta técnica toma como insumo al modelo entrenado \hat{f} y un set de datos X para devolver como resultado una matriz Φ donde cada elemento Φ_{ij} indica la forma en que el predictor j influye en la predicción que \hat{f} estima para i en relación al promedio de las predicciones \hat{f} en X .²⁴ De esta manera, se pueden obtener dos *outputs* a partir de SHAP. Por un lado la importancia de cada predictor, al computar $\sum_i |\Phi_{ij}|$ para cada una de las variables j . Por otro lado, si se grafican los valores de Φ_{ij} con los de X_{ij} se obtienen gráficos de dependencias SHAP, que pueden visualizar

²¹ Más detalles del proceso de optimización se puede consultar en [XGBoost](#).

²² Como se mostrará en los resultados, la técnica calcula la importancia que tiene cada variable en la predicción final, sin embargo, la forma en que esa variable genera dicho nivel de importancia resulta en cierto punto una “caja negra”.

²³ Este análisis se realiza siguiendo a [Soules \(2020\)](#) y [Molnar \(2022\)](#).

²⁴ Valores absolutos altos de Φ_{ij} sugieren un alto impacto de j en la predicción de i (independientemente de la dirección de la influencia), y viceversa, valores cercanos a cero sugieren bajo impacto de j en i .

la forma de impacto del predictor j en la predicción del modelo.²⁵ En el presente trabajo se hace uso de ambos outputs.

Adicionalmente, se agrega como medida de importancia relativa de variables el estadístico de ganancia (*gain*), que calcula el aporte total de cada variable a la predicción a partir de cada ramificación generada en los árboles de clasificación.²⁶

3.2.4 Métrica de evaluación: AUC-ROC

La curva ROC se ha introducido recientemente para evaluar los algoritmos de ML, al comparar el desempeño predictivo en el rango completo de las distribuciones de clases y los costos de error (Ling et al., 2003). Así, la misma muestra el desempeño de un modelo de clasificación para todos los umbrales o puntos de cortes,²⁷ graficando dos parámetros

- Tasa de Verdaderos Positivo (TPR) o *sensibilidad*
- Tasa de Falsos Positivos (FPR) o $1 - \textit{especificidad}$

TPR (sinónimo también de *recall*) es definido como

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

Mientras que FPR es

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

Luego, la curva ROC grafica TPR vs. FPR en diferentes umbrales de clasificación, mostrando así la habilidad que presenta el modelo para separar las clases.²⁸ Sin embargo, evaluar el desempeño de un estimador tantas veces como umbrales de decisión existan resulta una tarea ineficiente. Afortunadamente, el área bajo la curva ROC (AUC-ROC) puede solucionar este inconveniente.

²⁵ Por supuesto, se aclara que toda relación encontrada no pretende de ninguna forma tener una interpretabilidad causal. Sino más bien la técnica profundiza la manera en que cada variable se correlaciona con el *outcome*.

²⁶ Para una explicación más detallada se puede ver [XGBoost](#).

²⁷ Los modelos estiman probabilidades, por lo que para clasificar un resultado dentro de una u otra clase se debe establecer un umbral de decisión, comúnmente se elige 0.5. Sin embargo, este umbral puede no generar resultados satisfactorios cuando las clases de la variable dependiente se encuentran desbalanceadas.

²⁸ El desarrollo de este apartado se realiza con base en los análisis de [Ling et al. \(2003\)](#), [Hand y Till \(2001\)](#), como también, el análisis de [Google](#) y [Brefeld y Scheffer \(2005\)](#).

AUC-ROC mide toda el área por debajo de la curva, por lo que provee una medida agregada del desempeño entre todos los posibles umbrales de clasificación. Siguiendo a [Brefeld y Scheffer \(2005\)](#), sea f una función de decisión, la métrica representa la chance de que f asigne una probabilidad más alta a una observación positiva aleatoriamente seleccionada (y^+ que sería la salida del mercado laboral en nuestro caso) que a una observación negativa aleatoriamente seleccionada (y^- o no salida), esto es

$$AUC-ROC(f) = Pr(f(y^+) > f(y^-)) \quad (11)$$

Si bien $AUC-ROC$ en la Ecuación 11 refiere a la verdadera distribución de las clases, la misma puede ser estimada a través de la siguiente Ecuación 12. Dado los eventos positivos y negativos

$$\widehat{AUC-ROC}(f) = \frac{\sum_{i=1}^{n^+} \sum_{j=1}^{n^-} 1_{f(y^+) > f(y^-)}}{n^+ n^-} \quad (12)$$

La doble sumatoria en la Ecuación 12 itera sobre todos los pares posibles de positivos y negativos. Cada par que satisface $f(y^+) > f(y^-)$ contribuye con $\frac{1}{n^+ n^-}$ al valor total $\widehat{AUC-ROC}(f)$.

AUC-ROC es deseable por las siguientes dos razones:

- Invarianza a la escala. La métrica mide qué tan bien se clasifican las predicciones, en lugar de sus valores absolutos.
- Invarianza al umbral de clasificación. Mide la calidad de las predicciones del modelo independientemente del umbral de decisión de clasificación elegido.

El intervalo de dicha métrica es $[0,1]$, donde a mayor AUC-ROC implica mejor desempeño predictivo, mientras que un valor igual a 1 representa una separación perfecta de clases. En cambio, un valor igual a 0.5 implica que el modelo predice de la misma forma que un modelo totalmente aleatorio, por lo que un valor menor a 0.5 implica un desempeño peor que el azar.

Los cálculos de AUC-ROC se computan con sus respectivos intervalos de confianza al 95% a través de bootstrap con 1000 repeticiones. A su vez, para comparar el desempeño de cada estimador se utiliza el test de [DeLong et al. \(1988\)](#) que determina si las diferencias de predicción existentes entre modelos son estadísticamente significativas.

3.2.5 Enfoque de validación

El objetivo dentro de un ejercicio de predicción es desarrollar modelos con capacidad de predecir correctamente datos desconocidos, por lo que es indispensable contar con conjuntos de validación y testeo (James et al., 2013; Varian, 2014; Mullainathan y Spiess, 2017). Siguiendo a Varian (2014), la predicción requiere de un set de entrenamiento o *training* para estimar el modelo, un subconjunto de validación para elegir el modelo, el cual no es utilizado directamente por el estimador para aprender los patrones que presentan los datos, y un subconjunto de testeo para evaluar cuán bien se desempeña el modelo elegido. Este último se utiliza para obtener una estimación final del desempeño del estimador ante datos desconocidos.²⁹ De esta forma, se optimizaría la predicción de datos nuevos o desconocidos, evitando caer en el sobreajuste. Para el presente trabajo se optó por establecer como grupo de validación a los datos correspondientes al año 2017. En el caso del grupo *test*, se utilizan los datos del 2018. Todas las observaciones restantes, es decir, que se generaron antes del año 2017 forman parte del grupo de entrenamiento. La distribución de observaciones se puede ver en el siguiente Cuadro

Tabla 1: Subconjuntos del dataset

Enfoque de validación		
Conjunto	Período	Observaciones
Entrenamiento	2004-2016	166243
Validación	2017	12354
Test	2018	12781

Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2018

3.3 Análisis exploratorio

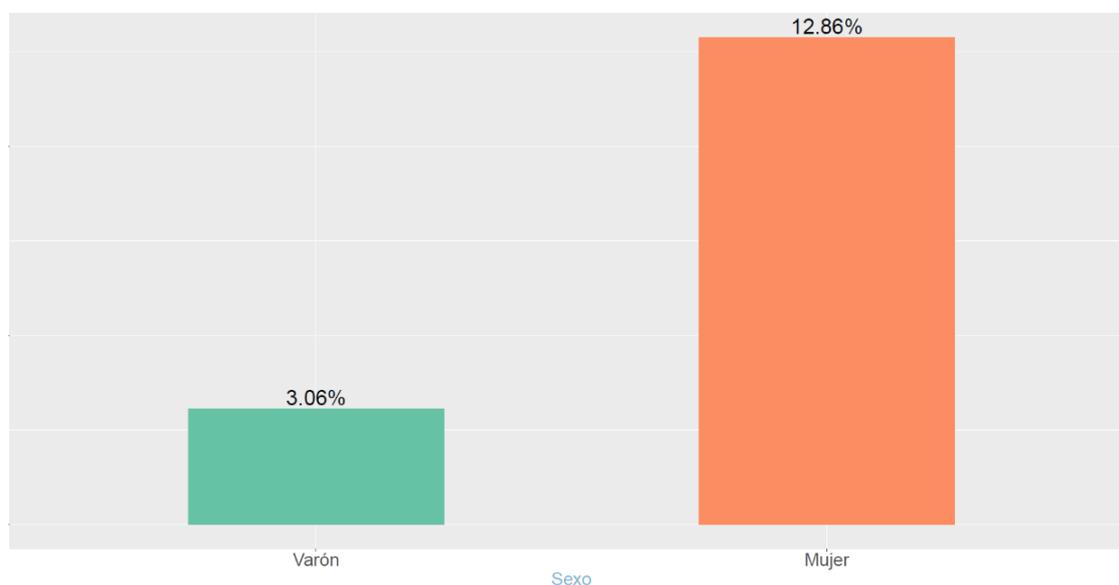
Debido a que el análisis de transiciones a la inactividad se centra en estudiar las diferencias de género a la hora de predecir dichas salidas, la etapa exploratoria hace foco en diferentes características para mujeres y varones por separado, utilizando el conjunto de entrenamiento.³⁰

²⁹ No está de más aclarar que para los tres subconjuntos se debe contar con los valores de las variables independientes, como también, estos deben conocer los valores del *outcome* con motivos de comparabilidad de resultados.

³⁰ Los datos de validación y de testeo se consideran datos “desconocidos” para el modelo. Por lo que no se considera dentro del análisis descriptivo de este apartado.

En primer lugar, se muestra el flujo de personas activas que pasan a la inactividad en el siguiente período (Figura 1). Como se puede ver, dichas proporciones son considerablemente menores a las proporciones de activos que permanecen en el mercado de trabajo.³¹ No obstante, el flujo de mujeres que pasan a la inactividad es considerablemente mayor a la de los varones. El 13% de las mujeres activas salen de la fuerza laboral en el siguiente año, mientras que en los varones este porcentaje es del 3%, existiendo así una diferencia de 10 puntos porcentuales entre ambos géneros. Como se explicará con mayor detalle en el Apartado 4.1, podría existir un efecto desincentivo, como también, un trasfondo cultural en la heterogeneidad de varones y mujeres.³²

Figura 1: Proporción de transiciones a la inactividad de mujeres y varones



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

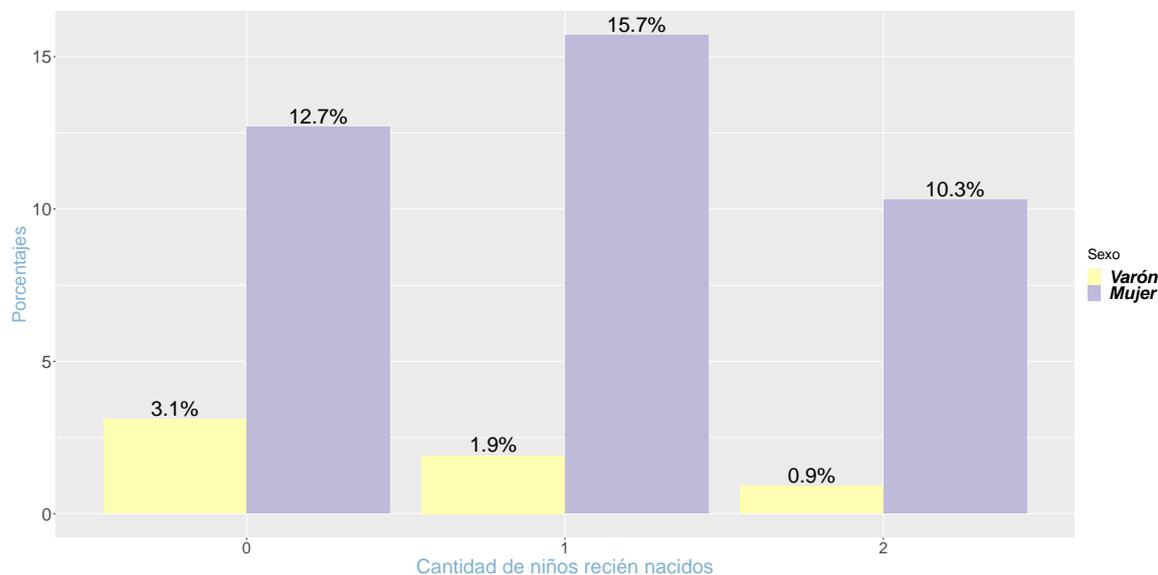
Las Figuras 2 y 3 muestran la proporción de salida según la presencia de recién nacidos en el hogar y el estado civil. En la primera se observa que, en hogares con presencia de un niño recién nacido la brecha de salida entre varones y mujeres aumenta en comparación a hogares sin recién nacidos, tanto por una mayor proporción de mujeres activas que salen de la fuerza laboral como por una menor proporción de varones activos que transita a la inactividad. Aunque, dicha proporción disminuye al tratarse de un hogar con dos niños recién nacidos. Adicionalmente, las brechas parecerían reforzarse al contraer matrimonio.

³¹ Esta distribución en la proporción de personas que transitan a la inactividad o se mantienen activos implica un fuerte desbalance en la variable a predecir, lo cual impediría generar resultados satisfactorios si utilizáramos como métrica de comparación de modelos la proporción de aciertos basado en un umbral de 0.5. Para estos casos de desbalance la métrica AUC-ROC toma mayor relevancia.

³² Se aclara que todas las estimaciones realizadas presentan una connotación exploratoria, las asociaciones encontradas no se interpretan como relaciones causales.

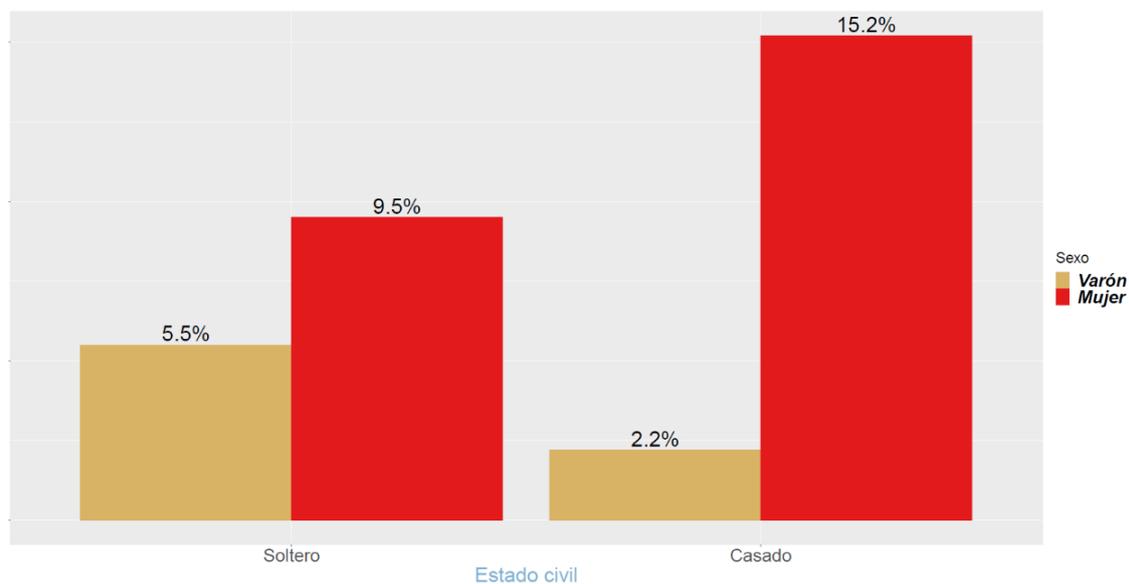
La Figura 3 sugiere que pasar del estado civil soltero a casado las probabilidades de salida en las mujeres aumentan, mientras disminuye la de los varones.

Figura 2: Proporción de transiciones a la inactividad de mujeres y varones según presencia de recién nacidos en el hogar



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 3: Proporción de transiciones a la inactividad de mujeres y varones según estado civil de la persona activa



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

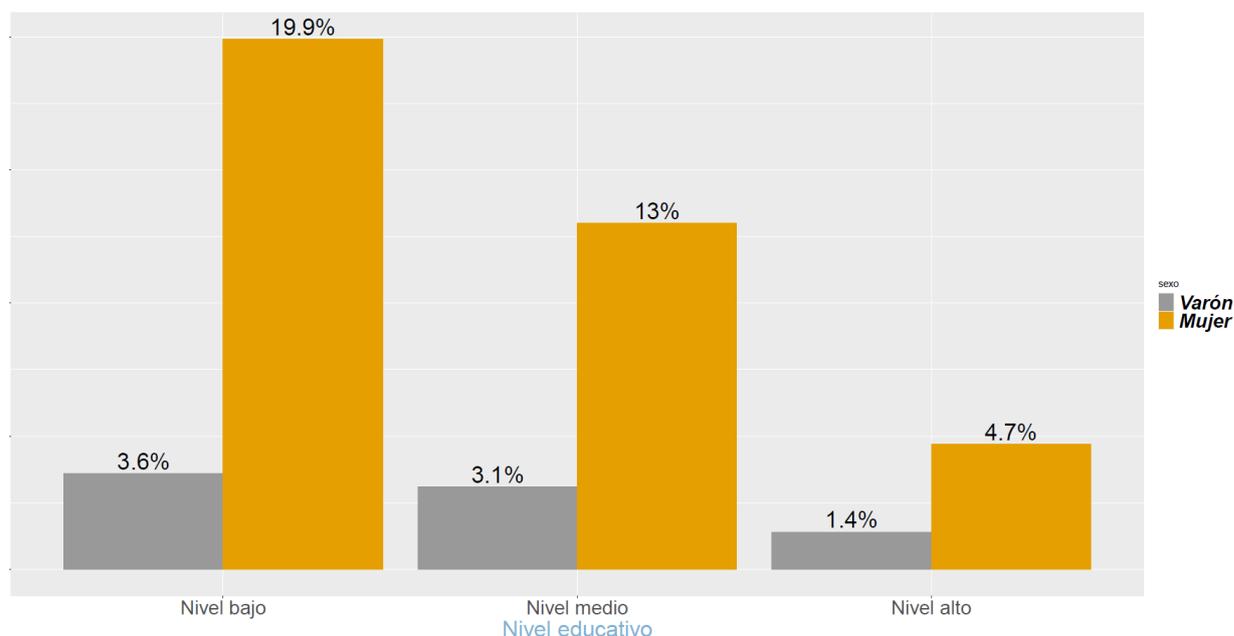
En la Figura 4 se muestran las proporciones de transición según la educación de cada individuo.³³ Como puede observarse, las proporciones de transición tienden a disminuir

³³ Los niveles educativos estándar de la EPH se han agregado en tres niveles, el nivel bajo corresponde a las personas con hasta secundaria incompleta, nivel medio hasta terciario incompleto y nivel alto

a medida que aumenta el nivel educativo, por lo que el gráfico sugiere una relación negativa entre este factor y la probabilidad de salir del mercado laboral. No obstante, las transiciones no son iguales por género. La probabilidad de salida de las mujeres cae casi 15 puntos porcentuales al pasar del nivel educativo bajo al alto, magnitud considerable no observada en los varones.

La brecha entre varones y mujeres es marcada entre y dentro de cada nivel educativo, donde la mayor diferencia se observa en el nivel educativo bajo, aquí el 4% de varones y 20% de mujeres activas salen de la fuerza laboral en el siguiente período, existiendo una brecha del 80%.³⁴ Por otra parte, la Figura 4 también sugiere que el efecto de la educación es mayor en la transición de las mujeres que la de los varones.

Figura 4: Proporción de transiciones a la inactividad de mujeres y varones según nivel educativo



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

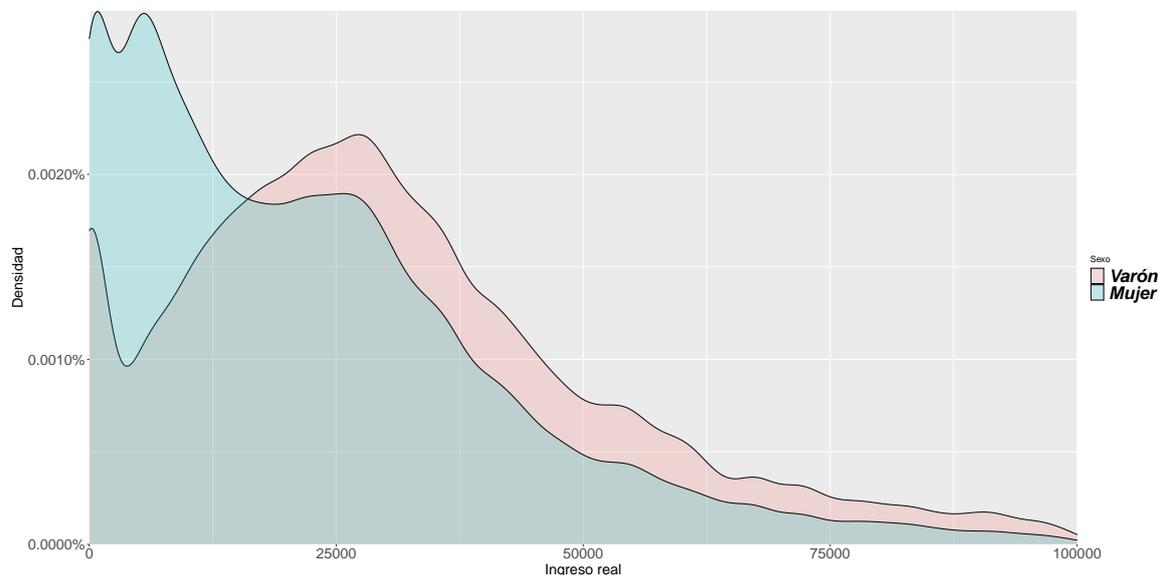
Por último, la Figura 5 muestra la distribución del ingreso de los ocupados que salieron de la fuerza laboral, tanto para mujeres como varones. En primer lugar, el gráfico sugiere una asociación negativa entre el ingreso y la probabilidad de salida en el siguiente período. La proporción de activos que pasan a la inactividad se acumula en los menores niveles de ingreso laboral. No obstante, la misma también muestra una considerable diferencia entre mujeres y varones. El flujo de salida de las trabajadoras aumenta en los niveles más bajos

corresponde a las personas activas con nivel terciario completo.

³⁴ Para llegar a esa brecha el cálculo es el siguiente: $1 - (ppv_{nb}/ppm_{nb})$, donde ppm_{nb} es la proporción de mujeres activas de nivel educativo bajo que pasa a la inactividad, y ppv_{nb} representa esa proporción para varones.

del ingreso, mientras que para los varones el pico de salida ocurre en niveles de mayor ingreso laboral en comparación a las mujeres.

Figura 5: Distribución del ingreso de la ocupación principal de los activos que salen del mercado laboral



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Las características relacionadas a la educación formal y al nivel de ingreso de la ocupación presentadas en esta sección se alinean con los resultados encontrados por la literatura de transiciones laborales. Se vislumbran mayores proporciones de salida a menores años de educación formal, como también, ante menores niveles de ingreso laboral,³⁵ por lo que dichos resultados exploratorios sugieren que las situaciones laborales como el nivel de ingreso o características de dotación personal, como el nivel educativo, podrían influir negativamente en la probabilidad de salida de la fuerza laboral. A su vez, las relaciones también se dan de manera heterogénea al diferenciar entre mujeres y varones activos. Las brechas de género persisten en los flujos de salida y parecen ser reforzadas por factores como los vistos en esta sección.

4 Resultados

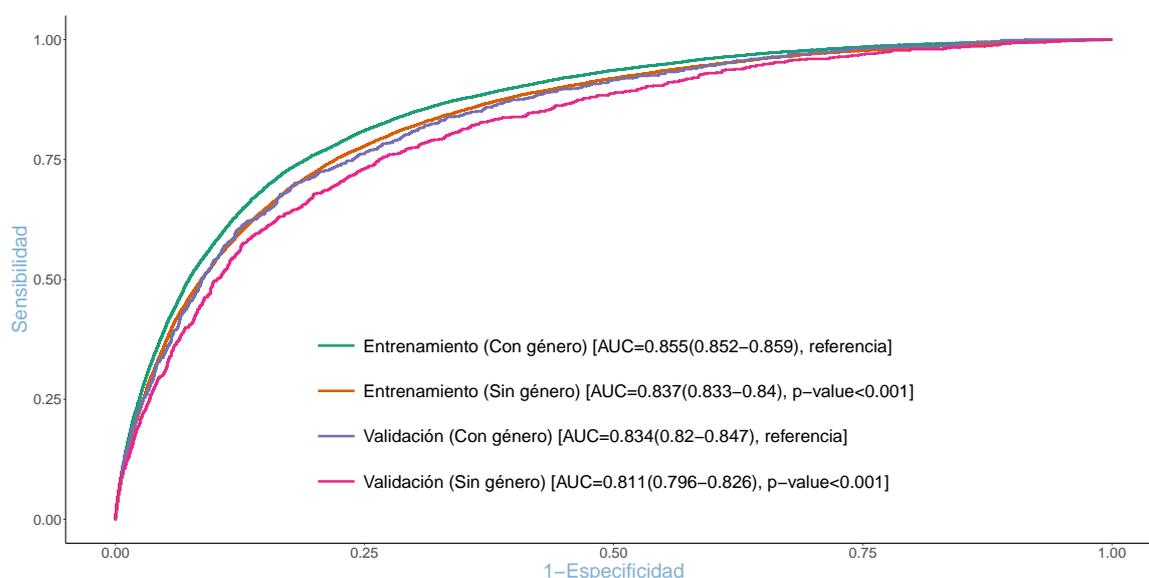
4.1 Modelos de predicción

³⁵ Si bien el trabajo presenta dos factores relacionados a cuestiones laborales y personales, en el Anexo C.1 se muestran otras características vinculadas a rasgos laborales de los individuos, como también, características personales que se analizaron previamente en la sección 2.

4.1.1 Regresión logit

La Figura 6 presenta la *performance* predictiva de las regresiones logit a través de la curva ROC y el cálculo del área debajo de la curva. Por un lado, se muestran los resultados de la curva AUC-ROC para una especificación con y sin incluir la *dummy* de género, para el grupo de entrenamiento. Luego, la *performance* para las mismas especificaciones en el grupo de validación.

Figura 6: Curvas ROC regresión logit. Desempeño predictivo en datos de entrenamiento y validación



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

En primer lugar, se destaca que la predicción de las transiciones incluyendo la *dummy* de género es significativamente mayor que la especificación sin considerarla. El aporte del género aumenta el poder predictivo aproximadamente 2 puntos porcentuales para ambos subconjuntos de datos.³⁶ Por supuesto, menores resultados en el grupo de validación para ambos modelos se corresponde con el sobreajuste que ocurre en datos que se usan para entrenar y predecir al mismo tiempo (grupo *train*), por lo que los resultados en datos nuevos (validación) son los que verdaderamente se intentan estimar. De esta forma, el mejor modelo logit tiene un 83% de chances de asignar un valor de probabilidad mayor a un evento de salida que a un evento de no salida.

Para ordenar la influencia relativa del género con las demás variables independientes se utiliza el método *forward stepwise selection*, que selecciona las variables que mayor ganan-

³⁶ Cabe aclarar que la especificación con género se corresponde con la inclusión de esta y su interacción con los demás predictores.

cia adicional en la predicción le agregan al modelo (James et al., 2013). El proceso de selección inicia desde una especificación sin predictores y agrega las variables, una a la vez, hasta llegar al modelo con el set completo de variables.³⁷ El criterio de selección se realiza en base al AUC-ROC sobre el grupo de validación, es decir, se ajusta el modelo con cada una de los predictores en el grupo de entrenamiento, pero se selecciona aquel que genere el mayor AUC-ROC adicional a la hora de predecir nuevos datos.³⁸

Los resultados se observan en la Figura 7 que muestra el ranking de las 18 variables principales seleccionadas para un modelo sin interacciones pero incluyendo la *dummy* de género.³⁹ Aquí se puede ver que la variable género se posiciona en el segundo lugar, después del logaritmo del ingreso laboral real. Así, el aumento adicional de 2 puntos al agregar dicho predictor al modelo (visto en la Figura 6) es el segundo en términos de influencia en la predicción. A su vez, la *dummy* de género presenta el segundo mayor aumento de ganancia adicional para predecir salidas del mercado laboral.

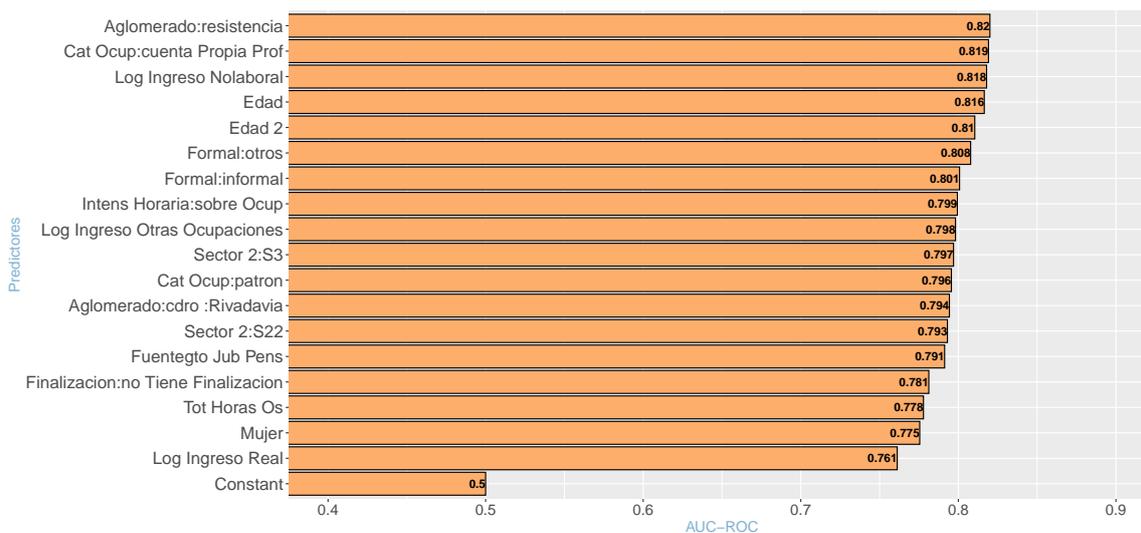
Dentro de las variables más relevantes para predecir salidas futuras a la inactividad las características personales y laborales tienen un rol importante. El ingreso y la cantidad de horas trabajadas en la ocupación secundaria son dos de las principales seleccionadas por el modelo logit, mientras que también selecciona la categoría que indica si el trabajo no tiene finalización. A su vez, se encuentran características como la edad y la formalidad, factores que según trabajos previos se asocian negativamente con la probabilidad de salida (Theodossiou y Zangelidis, 2009; Theodossiou, 2002; Theeuwes et al., 1990; Tasci y Tansel, 2005; Nilsen et al., 2000; Marston et al., 1976). Por otra parte, los sectores S3 y S22 se encuentran entre las principales seleccionadas, el primero representa el sector manufacturero mientras que en el segundo se encuentran los desempleados sin ocupación anterior. En ambos casos la correlación parcial indica una asociación negativa con la probabilidad de salir del mercado laboral (Tabla 4 del Anexo). Aunque en el caso de

³⁷ Previo a ejecutar el algoritmo de selección se aplica la técnica de *one-hot-encoding*, que transforma cada categoría de una variable *string* en una variable binaria con un 1 indicando si la observación tiene la categoría en cuestión.

³⁸ Usualmente, los estadísticos más utilizados para elegir aquella variable que tenga mayor influencia adicional son el criterio de información Akaike (AIC) o el bayesiano (BIC) para modelos de clasificación (R^2 para modelos de regresión). A su vez, el procedimiento usualmente utilizado ajusta y selecciona sobre el grupo de entrenamiento y la selección del modelo final se realiza evaluando esa especificación óptima en el grupo de validación (James et al., 2013). Como el objetivo aquí se centra en la selección de la variable, y no la de la especificación, se cree más adecuado modificar el procedimiento de la selección como lo indicado antes.

³⁹ Si bien la especificación con todas las variables es considerablemente más numerosa, se optó por agregar únicamente las primeras 18 debido a que las que siguen no aportan una ganancia adicional grande y el algoritmo tiene un tiempo de ejecución elevado, por lo que el costo beneficio de agregar más variables a la selección es alto. Esta decisión se aplica a todas las iteraciones de selección realizadas con logit.

Figura 7: Importancia de predictores. Método *forward stepwise selection*



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

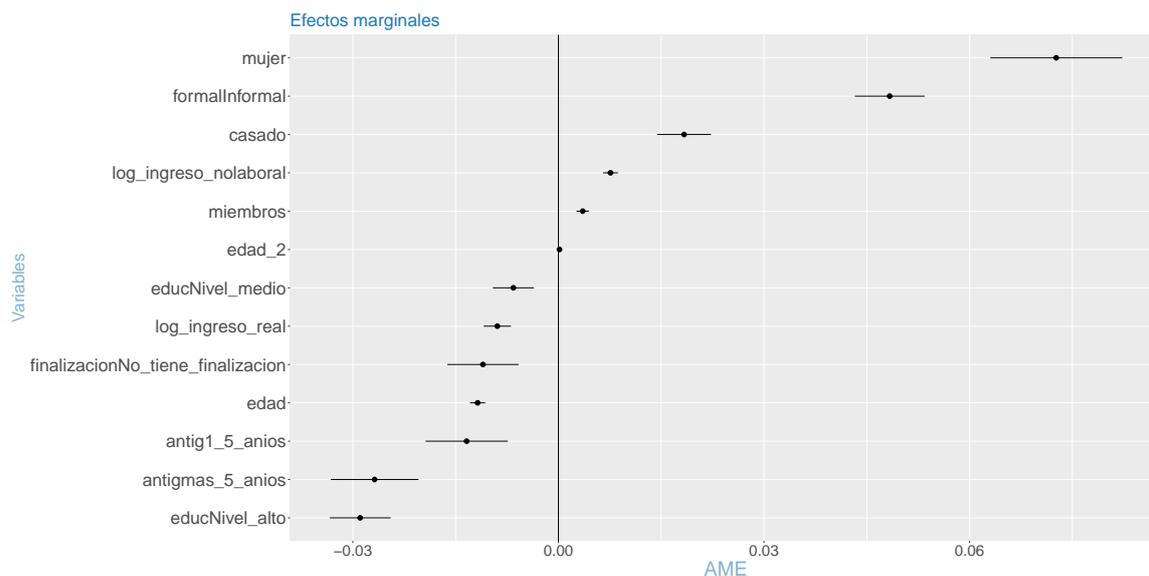
S22 la diferencia no es significativa. La relación negativa de S3 indica que un trabajador del sector manufacturero, o desempleado con experiencia previa en el sector, tiene menos chances de pasar a la inactividad en el siguiente período en comparación a un activo del sector agrícola (categoría de referencia). Quizás la diferencia significativa este asociado con la mayor presencia de puestos formales, que se relacionan con mayor estabilidad en el puesto, lo cual implica una mayor probabilidad de mantenerse en este.

Luego, se realiza el mismo procedimiento de selección agregando las interacciones de la *dummy* de género con los demás predictores (Figura 31 del Anexo C.2). Aquí se destaca la interacción entre el género y el índice líder, que aproxima los movimientos del ciclo económico nacional. Como también, la interacción del género con el nivel educativo alto y las horas trabajadas en la ocupación secundaria, que se encuentran entre las principales variables seleccionadas.

La Figura 8 muestra el efecto marginal del género en la predicción de salida junto con otras de las principales variables seleccionadas. Ser mujer aumenta las probabilidades estimadas de pasar a la inactividad en el siguiente período en casi 7 puntos porcentuales en relación a los varones. Un efecto de aproximadamente 6 puntos porcentuales mayor al que presenta el ingreso real de la ocupación principal y la edad, predictores que se encuentran entre los principales seleccionados en la Figura 7.

Como se explicó antes, la disparidad entre varones y mujeres ha sido documentado por la literatura (Theeuwes et al., 1990; Tasci y Tansel, 2005; Theodossiou y Zangelidis, 2009;

Figura 8: Regresión logit. Efectos marginales



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Hersch y Stratton, 1997; Marston et al., 1976; Tasci y Tansel, 2005; CEDLAS y BID, 2019).

Tasci y Tansel (2005) muestran que las mujeres ocupadas discontinúan su actividad laboral al perder su trabajo en mayor propensión que los varones, al igual que las desempleadas. Los autores explican que posiblemente en las mujeres exista un mayor efecto desincentivo que las lleve a pasar a la inactividad más que al desempleo ante la pérdida de su trabajo, como también, mayor desincentivo a mayor tiempo buscando un empleo.

Sin embargo, las condiciones culturales y sociales, que diferencian roles dentro del hogar, posiblemente también jueguen un papel en dicha brecha de género. La Tabla 2 muestra algunas de las correlaciones estimadas por el modelo logit con las interacciones entre el género y las demás independientes, lo que representaría la especificación donde todos los predictores son seleccionados (Tabla 5 del Anexo). En esta salida se destacan patrones que sugieren la presencia de roles tradicionales que influyen en la salida de la fuerza laboral. La presencia de niños menores de 10 años y la condición civil de casada se asocian negativamente con la probabilidad de pasar a la inactividad en el siguiente período, sin embargo, en las mujeres ocurre el efecto opuesto.⁴⁰

En cuanto al ingreso laboral, las estimaciones logit estiman correlaciones negativas entre

⁴⁰ Este patrón cultural se encuentra en trabajos anteriores. Hersch y Stratton (1997) muestran que las mujeres, especialmente las casadas, gastan el triple de su tiempo en actividades del hogar y están más preparadas para dejar el trabajo ante cuestiones relacionadas a la familia en comparación a los varones. Dicho comportamiento restringe la posibilidad en las mujeres de gozar de una vida laboral plena, como también, impide lograr una administración igualitaria entre trabajo y hogar para ambos géneros.

el nivel de ingreso laboral y la probabilidad de salida. Mientras que al interactuar con el género la diferencia no es significativamente distinto de cero. Así, tanto para varones como mujeres a mayor ingreso laboral menor es la probabilidad de salir de la fuerza laboral en el siguiente período. Resultado que coincide con lo documentado por la literatura.⁴¹

Los trabajos previos de transiciones laborales también encuentran evidencia acerca del nivel educativo y la probabilidad de salida. Como se argumenta en [Theodossiou y Zangelidis \(2009\)](#), mayores niveles educativos traen consigo mayores oportunidades laborales, lo cual se asocia con una menor probabilidad de salir del mercado de trabajo. En los resultados del presente trabajo, solamente el nivel educativo alto es estadísticamente significativo. Aunque al interactuar a esta última con la *dummy* de género la significatividad se da en el nivel medio. De todas formas, el signo siempre se mantiene, a mayor nivel educativo menores probabilidades de salir del mercado de trabajo.

Respecto a la informalidad, el efecto resulta ser el mismo para varones y mujeres. Encontrarse en condición de informalidad, que frecuentemente se vincula con relaciones laborales precarias e inestables, aumenta las probabilidades de salida de la fuerza laboral. En cuanto a las condiciones del entorno, la tasa de desempleo local no es estadísticamente significativa, aunque sí lo es su interacción con la *dummy* de género y su asociación es positiva, sugiriendo en las mujeres un patrón mencionado en la sección 2: encontrarse en un entorno (aglomerado en este caso) con alta tasa de desempleo aumenta las probabilidades estimadas de las mujeres de salir del mercado laboral.

⁴¹ En su trayectoria laboral, los individuos acumulan habilidades y conocimientos dentro del trabajo en forma de capital humano general o específico. Dicha acumulación tiene un efecto positivo en los ingresos. A su vez, el efecto de la inversión en capital humano sobre los ingresos depende en la longitud de tiempo general gastada en el empleo. Por lo tanto, personas que transitaron hacia la inactividad interrumpen la acumulación de capital humano y así su nivel de ingresos, lo cual sugiere que este último y salida pueden estar negativamente correlacionadas ([Theodossiou y Zangelidis, 2009](#)).

Tabla 2: Regresión logit

	log-odds
Miembros menores 10 a	-0.0919* (0.0475)
Mujer:Miembros menores 10 a	0.1481*** (0.0531)
Ingreso real (log)	-0.1350*** (0.0362)
Mujer:Ingreso real (log)	-0.0322 (0.0418)
Edad	-0.2292*** (0.0185)
Edad ²	0.0032*** (0.0002)
Mujer:Edad	0.0265 (0.0221)
Mujer:Edad ²	-0.0006** (0.0003)
Casado	-0.3403*** (0.0604)
Mujer:Casado	0.9073*** (0.0744)
Cant. miembros del hogar	0.0658*** (0.0147)
Nivel educativo medio	-0.0171 (0.0489)
Nivel educativo alto	-0.4771*** (0.1036)
Mujer:Nivel educativo medio	-0.1275** (0.0575)
Mujer:Nivel educativo alto	-0.1115 (0.1166)
Trabajador informal	0.5109*** (0.0797)
Tasa de desempleo local	0.7481 (0.9349)
Mujer:Trabajador informal	0.3964*** (0.0932)
Mujer:Tasa de desempleo local	-1.8306* (1.0756)
Ingreso no laboral (log)	0.3339*** (0.0183)
Mujer:Ingreso no laboral (log)	-0.2702*** (0.0216)
AIC	66803.0103
BIC	70150.0931
Log Likelihood	-33067.5051
Deviance	66135.0103
Num. obs.	166243

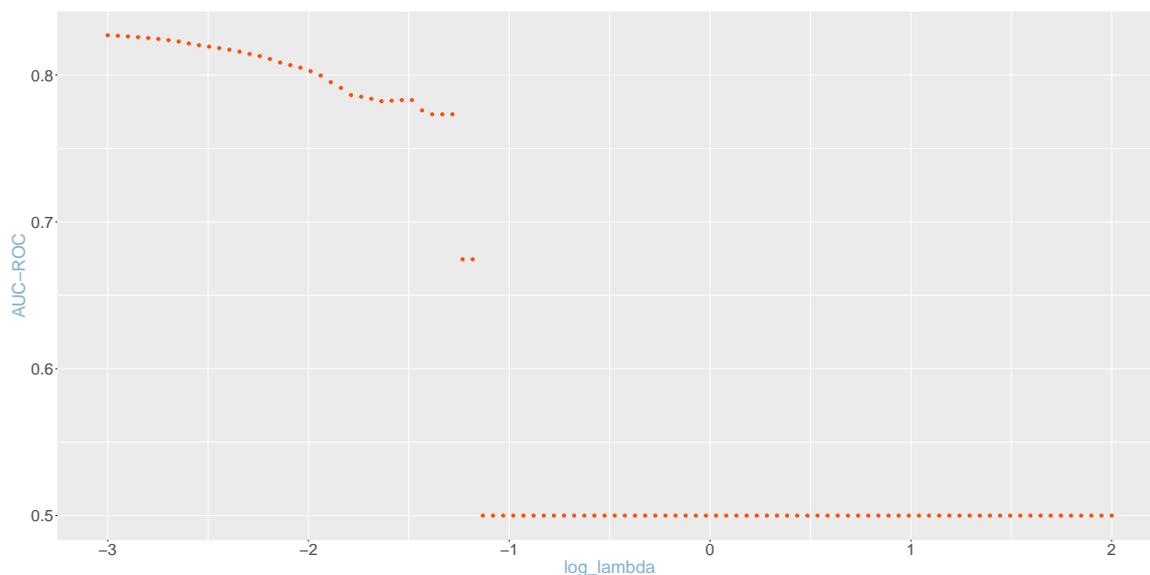
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

4.1.2 Regresión lasso

Para realizar la estimación lasso se ajustan diferentes regresiones optimizando la función objetivo de la Ecuación 5 para cada valor λ establecido previamente, utilizando como input el grupo de entrenamiento. Posteriormente, cada modelo es evaluado con la métrica AUC-ROC en el grupo de validación. Así, el modelo óptimo es aquel cuyo lambda genera la mejor *performance* predictiva, esto es, el mayor valor de AUC-ROC para los datos de validación.⁴²

Para el trabajo se establece una secuencia de 100 valores lambda que van desde 0.001 a 100. En la Figura 9 se muestra la evolución del AUC-ROC para cada valor de la grilla de lambdas (en logaritmo) utilizando la especificación con la *dummy* de género y la interacción entre esta y los demás predictores.

Figura 9: Regresión lasso. Performance predictiva en selección del hiperparámetro óptimo



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

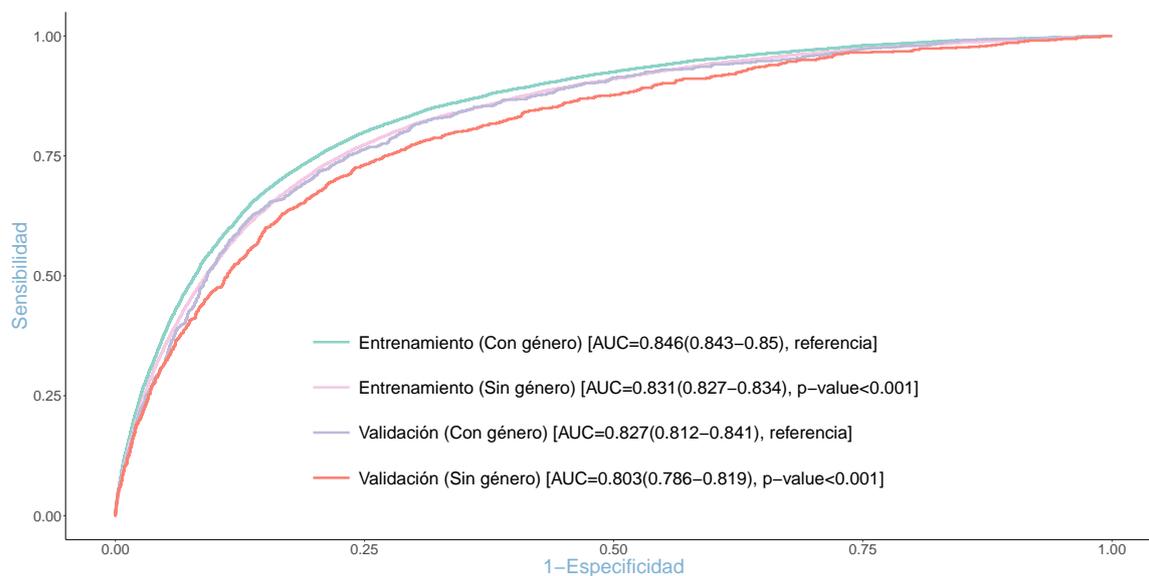
Como se puede observar, valores altos de lambda penalizan todas las variables predictoras y su desempeño predictivo es 0.5, que es equivalente a un resultado completamente azaroso. A medida que la regularización disminuye, el algoritmo comienza a seleccionar predictores y la predicción aumenta hasta llegar a su valor máximo 0.827.

El desempeño óptimo con lasso se logra con un hiperparámetro $\lambda^* = 0.001$ y la curva ROC de este modelo y su versión sin género pueden observarse en la Figura 10. Aquí lasso

⁴² El desarrollo de la regresión lasso se realizó principalmente con la librería *glmnet* de R desarrollado en Simon et al. (2011).

presenta un patrón similar a los resultados en logit, el factor género aporta una ganancia adicional significativa en la predicción, cuyo incremento es de aproximadamente 2 puntos porcentuales.

Figura 10: Curvas ROC regresión lasso. Desempeño predictivo en datos de entrenamiento y validación



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

Una de las ventajas que presenta lasso es la selección de variables. La regresión penaliza coeficientes, por lo que dado un hiperparámetro de regularización lo suficientemente grande lasso lleva a variables con coeficientes bajos hacia exactamente cero, mientras mantiene los de mayor magnitud. Lo que lo vuelve un seleccionador de variables en términos de importancia en la predicción. De esta forma, lasso brinda una alternativa para poder evaluar la importancia relativa del género en la predicción de salidas. Para ello, se corre un modelo similar al de la Figura 7, es decir, una especificación sin interacciones pero incluyendo la *dummy* del género.

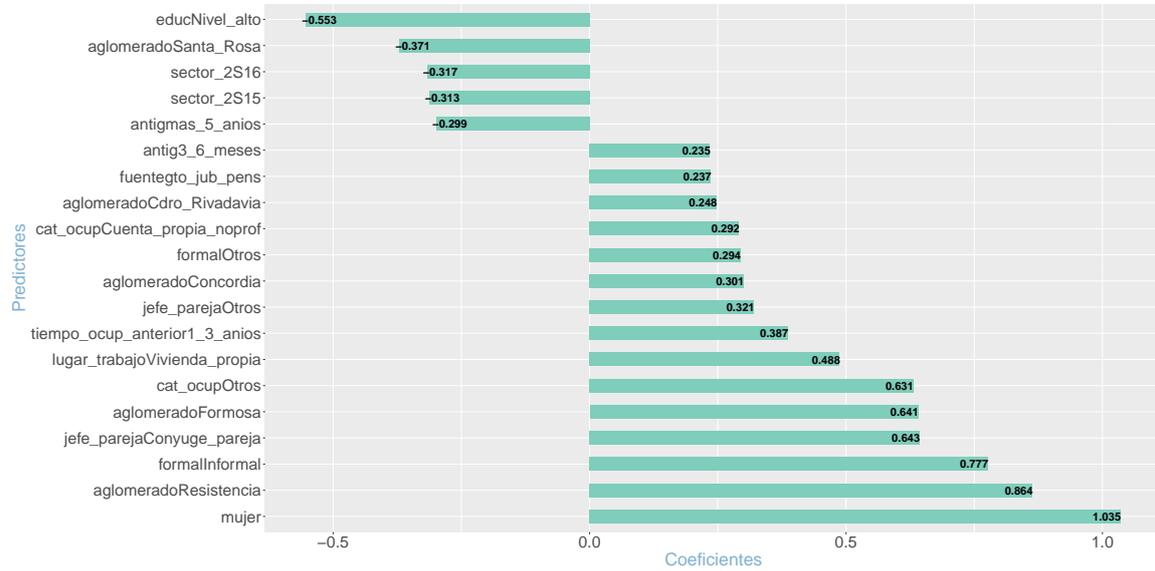
El modelo óptimo presenta el mismo valor óptimo $\lambda^* = 0.001$ y seleccionó 87 de 167 coeficientes disponibles.⁴³ Los 20 más importantes se observan en la Figura 11.

Coincidiendo con los resultados en logit, el factor género vuelve a encontrarse entre las principales, aunque en este caso se ubica en el primer lugar en términos absolutos. De la misma manera, la informalidad se mantiene como un factor relevante para la predicción. Ser mujer y/o encontrarse en el sector formal aumenta las probabilidades de salir del mercado laboral en el siguiente período en relación a los varones y/o trabajadores del

⁴³ Todos los coeficientes seleccionados se encuentran en la Tabla 6 del Anexo.

sector formal. Por otra parte, a diferencia de logit, lasso no selecciona al ingreso real laboral dentro de las más influyentes.

Figura 11: Principales variables seleccionadas regresión lasso



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

Resulta interesante notar que, la importancia del género se puede focalizar en algunos atributos al interactuar a este con los demás predictores (Tabla 7 del Anexo). Como se puede ver en la Tabla 7, las diferencias entre varones y mujeres interactúan con factores relacionados con la segmentación territorial y productiva (aglomerados y sectores productivos), y con otras características sociodemográficas como la relación de parentesco con el jefe de hogar y la situación conyugal de la mujer.

4.1.3 XGBoost

Para utilizar el modelo XGBoost se deben definir los valores de los hiperparámetros mencionados anteriormente, los cuales pueden afectar considerablemente el desempeño predictivo del estimador. Siguiendo a Soules (2020), se optó por utilizar el procedimiento *random search*, el cual consiste en definir un rango de posibles valores para cada hiperparámetro, de manera tal que el algoritmo realice la predicción para cada combinación de valores seleccionados aleatoriamente. Por lo tanto, habrá tantos modelos como combinación de hiperparámetros se especifique. El elegido es aquel cuya combinación genere el mayor AUC-ROC en los datos de validación (Bergstra y Bengio, 2012).⁴⁴ El rango de valores se

⁴⁴ El desarrollo del modelo XGBoost se realizó principalmente con la librería *xgboost* de R desarrollado en Chen et al. (2021).

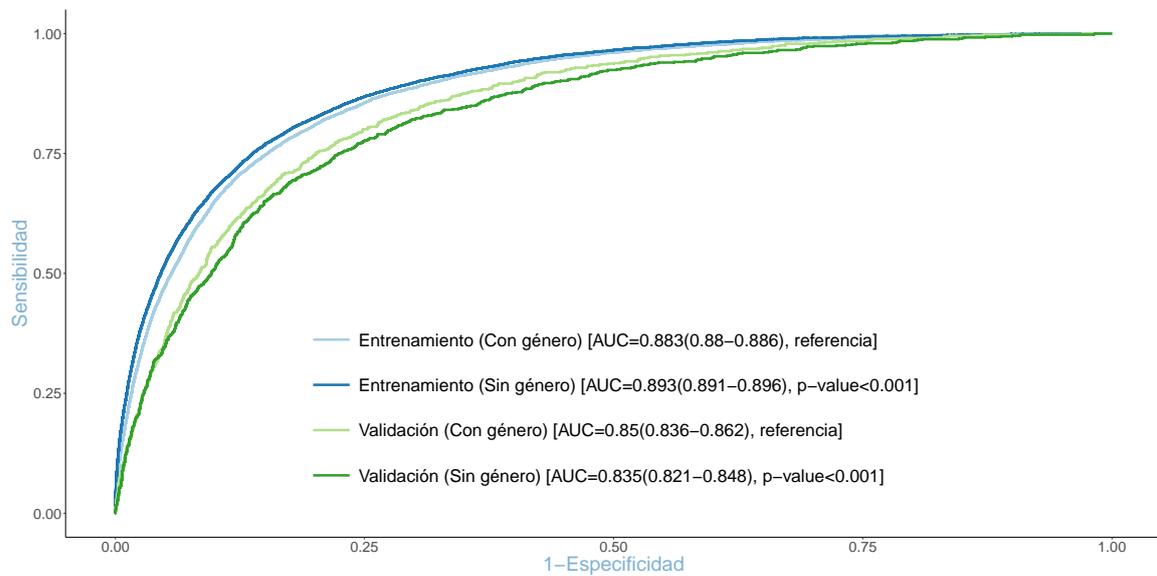
puede ver en la Tabla 3

Tabla 3: Performance predictiva

Hiperparámetro	Rango
size	100
nrounds	[200;600]
max_depth	[2;15]
eta	[0.001;0.01]
gamma	[0;1]
coldsample_bytree	[0.6;1]
min_child_weight	[1;20]
subsample	[0.75; 0.9]

Una vez estimado XGBoost utilizando *random search* se selecciona el mejor modelo. La Figura 12 muestra las curvas ROC para el modelo que incluye la *dummy* de género y la especificación sin incluirla.

Figura 12: Curvas ROC XGBoost. Desempeño predictivo en datos de entrenamiento y validación

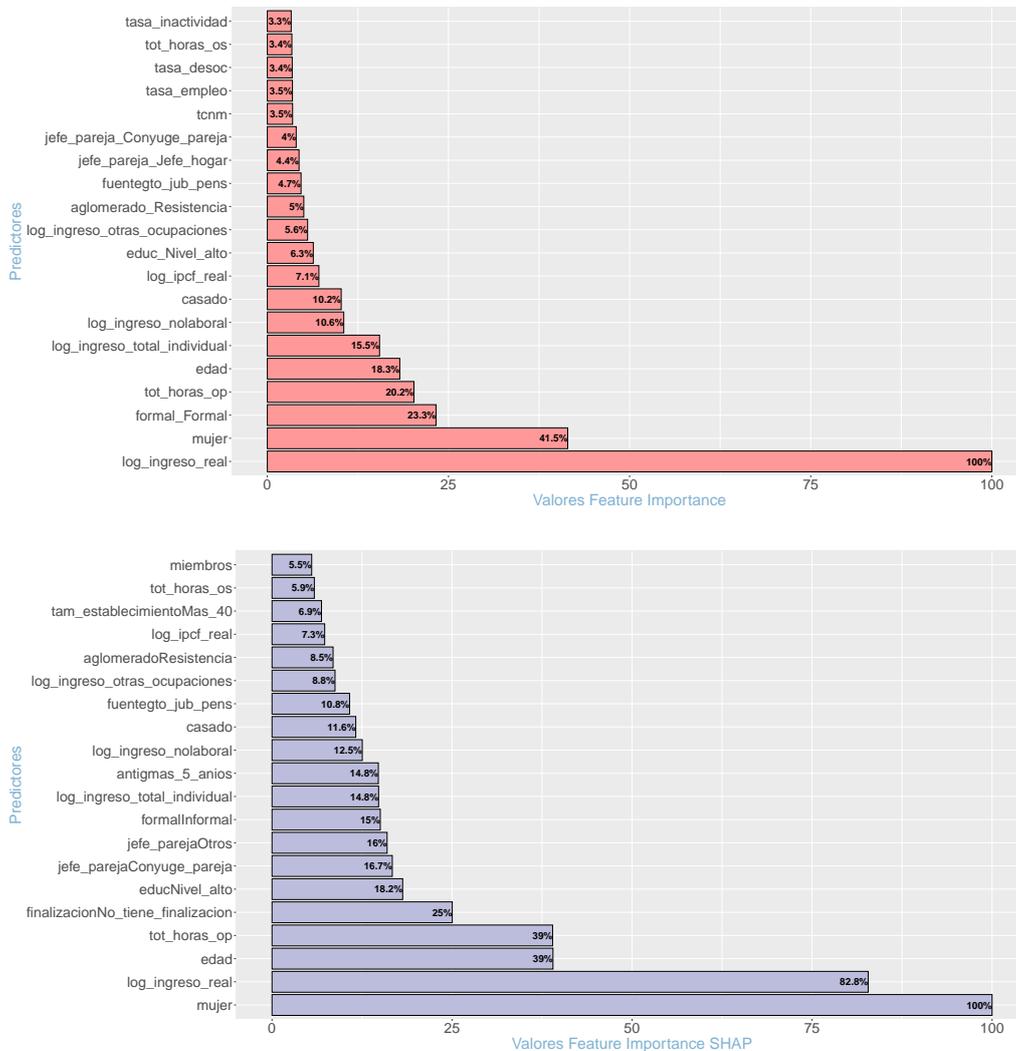


Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

Como se puede ver, tanto en el conjunto de validación como de entrenamiento se observan diferencias significativas al pasar de un modelo sin género a uno incluyendo dicho factor. El modelo XGBoost con género tiene un 85% de chances de que si se eligen dos observaciones al azar, una de “Salida laboral” y otra de “No Salida”, la probabilidad predicha para la “Salida” sea mayor que para la “No salida”, aumentando el poder predictivo en 2 puntos porcentuales en comparación al modelo XGBoost sin el factor género.

Por último, utilizamos el mismo modelo para computar la importancia de cada predictor bajo dos métodos: Gain y SHAP. Los resultados se encuentran en la Figura 13, que muestra la importancia de cada variable en relación al predictor que se ubicó en el primer lugar.

Figura 13: Importancia de atributos. Métodos Gain y SHAP

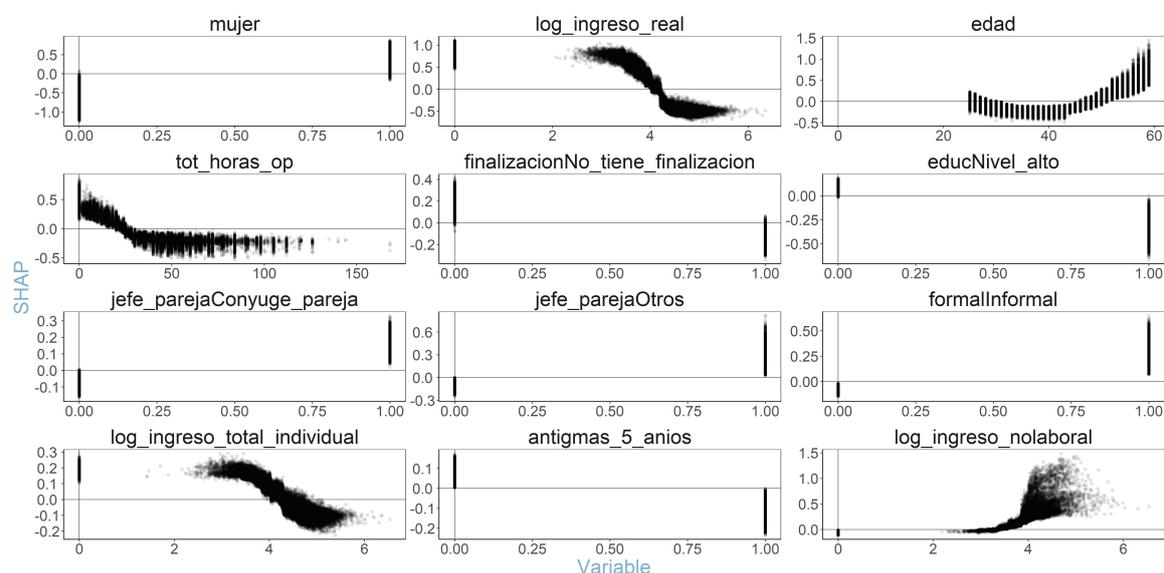


Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

Nuevamente, el género vuelve a ubicarse entre los primeros lugares. En el primer plot la influencia del género representa el 42% de la importancia del ingreso real laboral, mientras que en el segundo es el predictor más importante, estableciendo a este como variable de referencia. Estos resultados refuerzan lo encontrado previamente. La brecha entre varones y mujeres resulta fundamental para predecir salidas del mercado laboral, la misma es igual o más importante que el ingreso laboral. Por otro lado, ambas figuras comparten atributos como la educación, el estado civil, variables de ingreso, informalidad, entre otros.

Por último, para caracterizar la forma de utilización de las variables por parte de XGBoost se presentan los gráficos de dependencia SHAP. En la Figura 14 aparecen las 10 variables más importantes. A partir de cada gráfico se pueden extraer algunos rasgos que permiten entender mejor los patrones de movilidad laboral de la actividad hacia la inactividad. En primer lugar, y como analizamos previamente, ser mujer aumenta las probabilidades de salir del mercado laboral en el siguiente período, como también, encontrarse en la informalidad o no ser el jefe de hogar. A su vez, se encuentran características laborales que muestran un comportamiento esperado. A mayor antigüedad en el puesto, cantidad de horas de trabajadas e ingresos laborales menor es la probabilidad de salida.

Figura 14: Gráficos de dependencia SHAP. Atributos más importantes (Top 12)



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

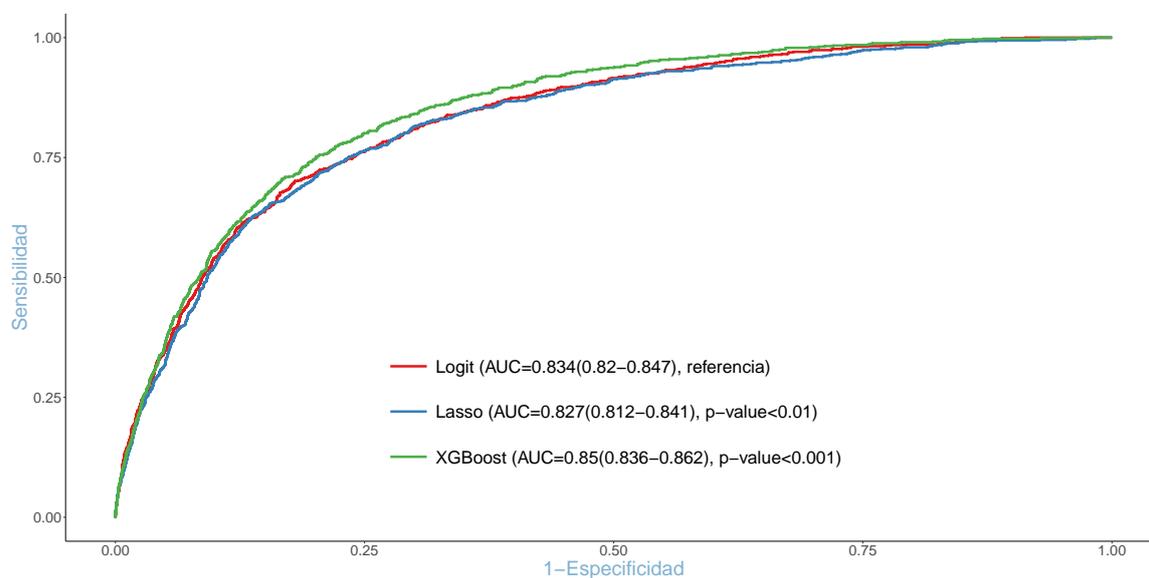
En el caso de la edad, esta pareciera tener forma de U, ya que a edades cercanas a los 25 años, que generalmente se consideran los primeros años en el mercado laboral poseducación superior, la probabilidad pareciera ser débilmente positiva. En cambio, a mayores edades esta comienza a disminuir hasta llegar a los 45 años aproximadamente, que es donde la probabilidad de salida comienza a aumentar. Por último, en lo que respecta a nivel educativo, el modelo encuentra un patrón relevante dentro del nivel alto, que correspondería a educación superior completa. Así, activos con nivel educativo alto tienen menos probabilidades de salir del mercado laboral en relación a activos con nivel educativo bajo.

4.1.4 Comparación de modelos

En los apartados anteriores se analizaron tres de los modelos utilizados para predecir salidas a la inactividad por separado. El primero fue el modelo logit, utilizado tradicionalmente por la literatura de movilidad laboral. En segundo lugar, el modelo lasso que agrega un hiperparámetro de penalización a la función objetivo de la regresión logística, con el objetivo de mejorar la predicción reduciendo la varianza de la estimación a costa de un aumento del sesgo. Tercero, el modelo XGBoost, que se ha vuelto uno de los estimadores más populares y utilizados a la hora de realizar ejercicios predictivos. En este apartado, se compara la *performance* predictiva de los tres modelos utilizados.⁴⁵

La Figura 15 muestra la curva ROC de cada modelo que incluye la variable género y las interacciones para los datos de validación. En primer lugar, la gráfica muestra la diferencia significativa que existe entre XGBoost y logit, dado que el primero tiene un 85% de probabilidades de asignar una probabilidad más alta a “Salidas” respecto a la “No salida”, lo que equivale a casi 2 puntos porcentuales más que el estimador logístico. Por otra parte, no se produce una ganancia en la predicción al pasar de la regresión logit a lasso, la diferencia entre ambas a favor del primero es estadísticamente significativa.

Figura 15: Curvas ROC. Comparación de desempeño predictivo en datos de validación



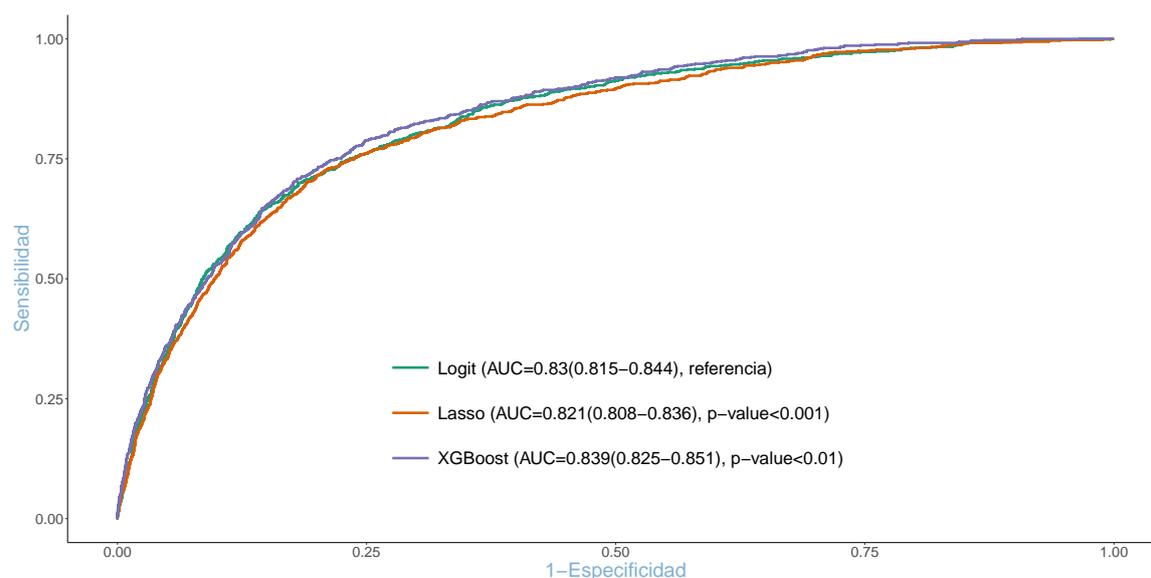
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2017

Por último, se realiza el mismo análisis para un tercer set de datos, el grupo de testeo, que

⁴⁵ Para este apartado se utilizaron las mejores versiones de cada estimador, es decir, las que contienen la variable de género con las interacciones entre esta y los demás predictores.

corresponde a datos del año 2018 que intentan predecir tanto salidas correspondientes a finales del mismo año como a movimientos del 2019. Los resultados para los tres modelos pueden verse en la Figura 16. Nuevamente aquí, la diferencia en la predicción entre XGBoost y logit es estadísticamente significativa, aunque la brecha entre ambos disminuye en comparación a lo visto en la Figura 15. Lo opuesto ocurre entre lasso y logit, la brecha entre ambos se amplía y se vuelve aún más significativa que al utilizar el set de validación.

Figura 16: Curvas ROC. Comparación de desempeño predictivo en datos de testeo



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2018

Vistos los tres modelos se destacan tres puntos importantes. En primer lugar, pasar de un modelo tradicional a uno más sofisticado aumenta el poder predictivo significativamente. Sin embargo, se destaca la competitividad del modelo logit que mantiene niveles de predicción cercano a un modelo más sofisticado como lo es XGBoost. Lo cual lo vuelve atractivo, debido a su simplicidad a la hora de aplicarlo. En segundo lugar, se destaca la importancia del factor género como predictor para estimar transiciones laborales futuras de corto plazo de la actividad a la inactividad. Incluir la *dummy* de género es al menos tan importante como pasar de un modelo simple a uno más sofisticado, e igual o más importante que una variable característica del mercado de trabajo, el ingreso laboral. Por último, coincidiendo con los resultados de [Salganik et al. \(2020\)](#), se puede observar un error de predicción considerable en los tres estimadores, lo que sugiere que dicho error podría estar fuertemente relacionado con la dificultad de predecir patrones sociales y débilmente asociado con la técnica usada para generar la estimación.

4.1.5 Umbral de decisión

Uno de los principales inconvenientes a la hora de predecir salidas de corto plazo de la fuerza laboral es la gran propensión de los activos de mantenerse en la mercado de trabajo, en comparación con el flujo de personas que salen de este. De hecho, observando la Figura 1 analizada anteriormente se puede ver que del total de personas activas solo el 16% transita hacia la inactividad en el siguiente período, por lo que la tarea se basa en un problema de predicción de clases desbalanceadas.

Como se dijo antes, ante este panorama el área bajo la curva ROC resulta una métrica ideal para predecir salidas del mundo laboral, debido a que analiza la *performance* predictiva para todos los umbrales de decisión, lo que lo hace independiente de cualquier punto de corte específico y de la presencia de desbalance en las clases. No obstante, en pos de contribuir en la elaboración de políticas para anticiparse y evitar probables salidas no deseadas, sería de gran ayuda utilizar el modelo de predicción para poder determinar cuántas personas saldrán en el siguiente período, algo que no resulta tan claro de ver al computar valores AUC-ROC.

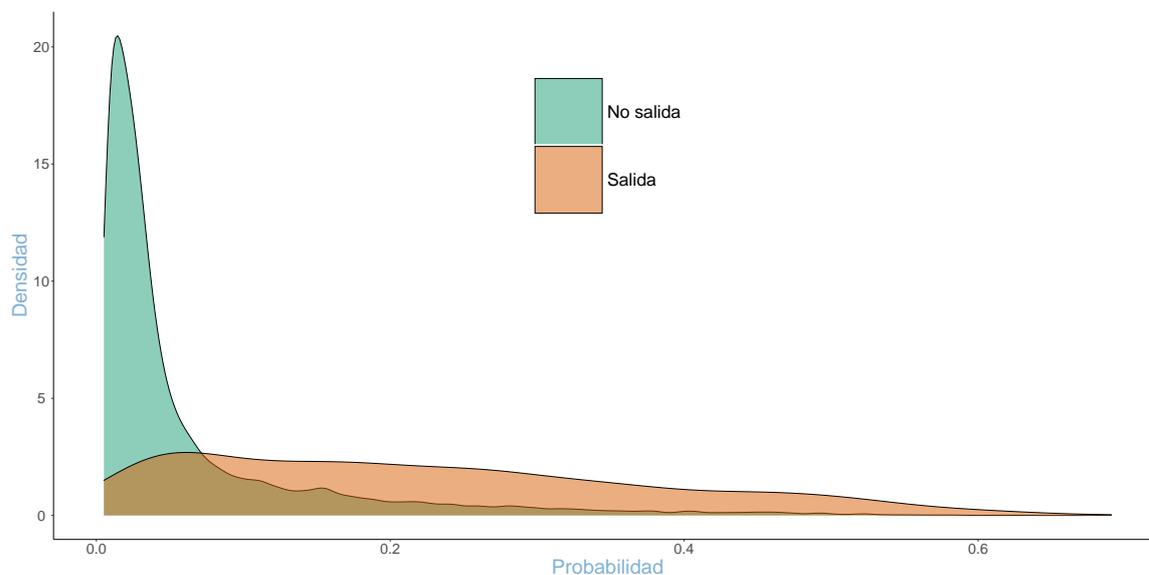
La Figura 17 muestra la distribución de probabilidad del modelo XGboost que mejor desempeño presenta entre los utilizados en este trabajo, esto es, la especificación incluyendo la *dummy* de género con las interacciones entre esta y los demás predictores, evaluado en los datos de validación.⁴⁶ Como se puede ver, si bien existe una superposición entre clases, el modelo puede separar ambas clases con un alto porcentaje de acierto.⁴⁷ Gran parte de las observaciones se acumulan en la no salida laboral, algo que el modelo XGBoost parece poder identificar, asignando una probabilidad predicha cercana a cero para la mayoría de estas observaciones.

A partir de la Figura 17 sería ideal determinar un umbral que, optimizando la predicción, pueda dar información precisa de cuántas personas saldrán del mercado laboral en el próximo período. Para comenzar, se utilizan diferentes umbrales de decisión y se computan la proporción de aciertos de la matriz de confusión resultante, lo que resulta en la medida conocida como *accuracy*. Los resultados se muestran en la Figura 18. El nivel de *accuracy*

⁴⁶ A efectos de no duplicar resultados se optó por utilizar los datos de validación, no así los de testeo. Aunque cabe aclarar que los resultados son similares.

⁴⁷ Una *performance* perfecta sería aquella donde la distribución de probabilidades de ambas clases no se superpongan en ningún punto, por lo cual existiría una separación precisa en la predicción de cada clase y obtendríamos como resultado un $AUC = 1$.

Figura 17: Modelo XGBoost con género y sus interacciones. Distribución de probabilidades del estimador



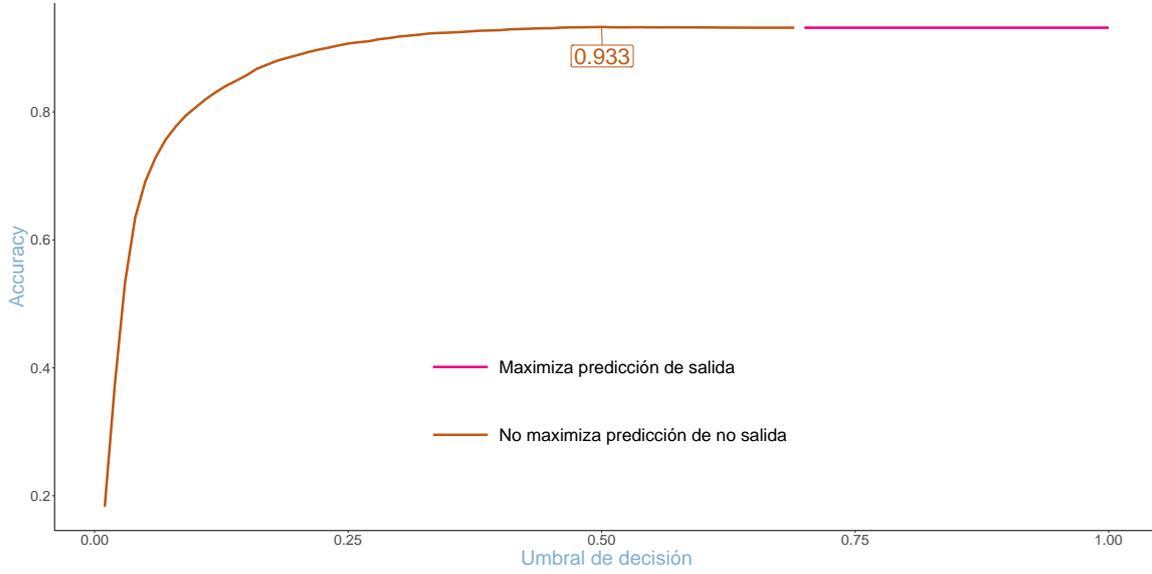
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2017

aumenta hasta llegar a un nivel estable que se empieza a dar en un umbral de decisión de aproximadamente 0.4. Mientras que el valor óptimo se genera justo en la mitad, es decir, bajo un umbral de decisión de 0.5. No obstante, el problema que presenta la medida de *accuracy* es la igualdad de ponderación para ambas clases.

En la Figura 18 se divide la evolución del acierto en dos grupos, ambos hacen referencia a si se maximiza o no la predicción de salida, es decir, pasado cierto umbral el modelo pasa a determinar todos los predichos como “No salida”, lo cual se comienza dar a partir del umbral 0.70 aproximadamente. Antes de llegar a este punto, la predicción también clasifica la “Salida”. De todas formas, a partir del umbral 0.40 aproximadamente la regla de decisión establece una clasificación muy cercana a fijar toda predicción como “No salida”. El problema detrás de *accuracy* es que supone previamente la misma ponderación para ambas clases a predecir, por lo que la medida supone que el costo de no salir del mercado laboral es igual a salir de este. Suposición desacertada debido a que, como se analizó antes, ambos escenarios no conllevan el mismo costo, la salida de la fuerza laboral representa un mayor costo que la no salida y resulta de mayor interés poder identificar correctamente este evento. Por lo que la maximización de *accuracy* llevaría a conclusiones erróneas.⁴⁸

⁴⁸ Idealmente, se podría pensar en calcular una matriz de costos que pondere el costo social de que se produzcan cada una de las clases, sin embargo, dicho cálculo escapa del alcance de este trabajo.

Figura 18: Modelo XGBoost con género y sus interacciones. Evolución métrica *accuracy*



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2017

Por lo tanto, para considerar clases desbalanceadas se utiliza el *F1-score*.⁴⁹ Para construir el indicador se retoma la definición de *Recall* vista en la Ecuación 9

$$Recall(TPR) = \frac{TP}{TP + FN} \quad (13)$$

Y se agrega la Ecuación de *Precision*

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

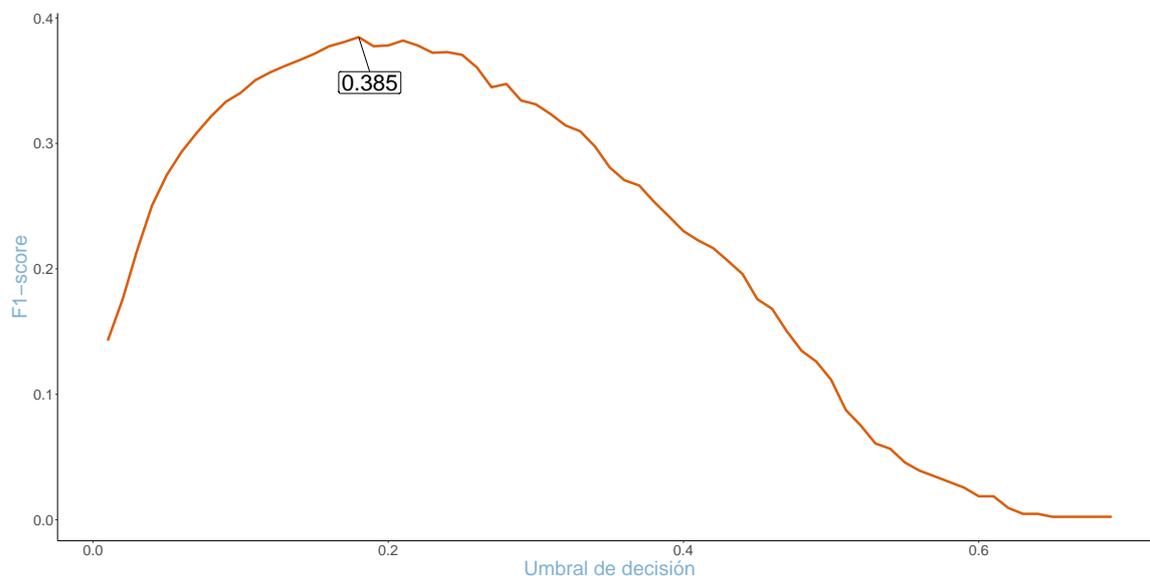
Recall es el ratio de observaciones positivas predichas correctamente (predecir salidas correctamente en este contexto) sobre el total de observaciones verdaderamente positivas, por lo que un alto *Recall* se genera con un bajo ratio de falsos negativos. Mientras que *Precision* es el ratio de observaciones predichas positivas correctamente sobre el total de observaciones que se predijeron como positivas. Por lo que una alta precisión se relaciona con generar una baja tasa de falsos positivos. A partir de ambos estadísticos se construye el *F1-score* de la siguiente manera

$$F1-score = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (15)$$

⁴⁹ Análisis siguiendo a [Khanna et al. \(2015\)](#).

Debido a que la Ecuación 15 toma en cuenta los resultados en los valores positivos, la métrica funciona mejor para contextos como el ejercicio de este trabajo, donde las clases se encuentran desbalanceadas. La evolución de la métrica *F1-score* se muestra en la Figura 19.

Figura 19: Modelo XGBoost con género y sus interacciones. Evolución métrica *F1-score*



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2017

Como se puede observar en el gráfico, la *performance* del indicador presenta forma de U invertida, aumentando en los primeros puntos de corte hasta un valor máximo, a partir del cual comienza a disminuir su desempeño, hasta llegar al umbral de 0.69 que es donde el modelo clasifica todas las observaciones como “No salida”, por lo que este punto representa el límite superior de *F1-score*. Claramente, la nueva métrica optimiza su valor en un umbral considerablemente menor que *accuracy*.

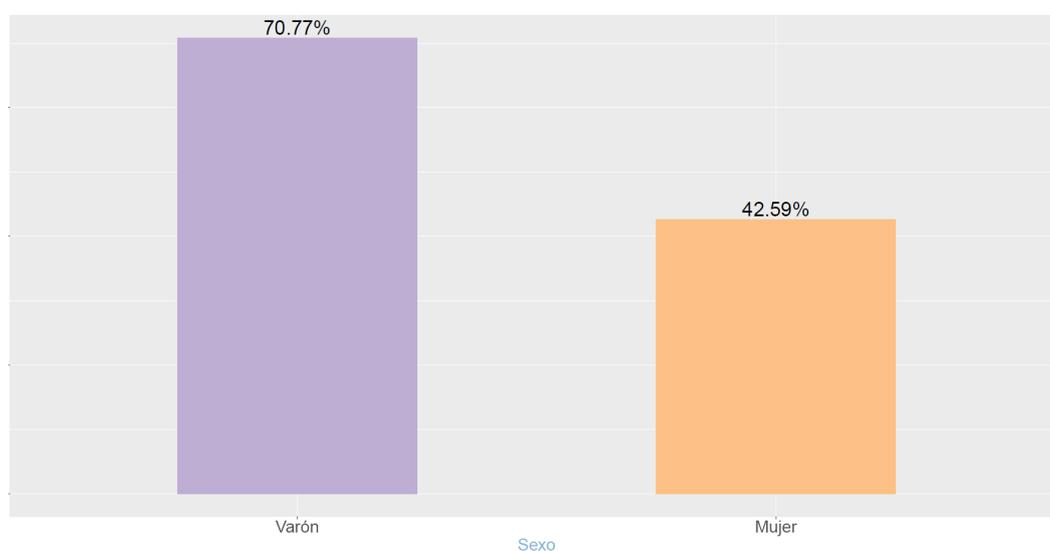
Por último, utilizando el umbral óptimo de 0.18 encontrado en *F1-score* el modelo proyecta una salida del 13%, es decir, del total de individuos activos durante el año 2017 el 13% pasará al estado de inactividad en el siguiente año. De los cuales 11 puntos porcentuales son mujeres.

4.1.6 Post-predicción: análisis de inserción laboral

En base a los resultados anteriores, las transiciones de corto plazo a la inactividad presentan significativas brechas de género. Las mujeres tienen mayores probabilidades de salir del mercado laboral en el siguiente período en relación a los varones, y cuya brecha se refuerza

al poner el foco en determinadas características personales y familiares de los activos. Las diferencias de género se magnifican en las mujeres casadas; cónyuges; perteneciente a hogares con al menos un recién nacido en el hogar; trabajadoras del sector informal; entre otras características. Sin embargo, el hecho de que los resultados de los modelos de predicción sean solamente de corto plazo, no es excluyente de que la problemática no pueda persistir en el tiempo. La Figura 20 muestra la proporción de mujeres y varones inactivos que se han insertado a la fuerza laboral en el siguiente período.⁵⁰

Figura 20: Proporción de mujeres y varones inactivos que ingresan al mercado laboral en el siguiente período



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Como se puede ver en la Figura 20 la probabilidad de volver al mercado laboral pareciera ser mayor para los varones, ya que del total de mujeres que pasaron de la actividad a la inactividad solamente el 43% retorna al mercado laboral en el siguiente período, lo que equivale a casi 30 puntos porcentuales menos que los varones. Resultado que, complementado a los resultados anteriores, no sólo sugiere que la movilidad de las mujeres es menor a la de los varones (como encuentra Theodossiou y Zangelidis (2009)), sino que el mayor flujo de salida de las mujeres podría ser permanente. Además, dicha brecha parece aumentar ante la presencia de menores de edad en el hogar; en personas casadas; jefes de hogar y en los deciles más bajos de la distribución del ingreso per cápita familiar (Figuras disponibles en el Anexo C.4).

⁵⁰ Con motivos de captar a aquellos inactivos que salieron del mercado en el corto plazo anterior al momento en que estos son encuestados, se toma solamente a aquellos inactivos que durante los 12 meses previos a la encuesta estuvieron trabajando o buscando trabajo. De esta manera, el análisis se realiza sobre 3523 inactivos.

5 Conclusiones

El presente trabajo se propuso estudiar la viabilidad de desarrollar modelos que predigan de la manera más certera posible la transición laboral de corto plazo en el mercado de trabajo argentino, enfocando el análisis en la predicción de la salida de la fuerza laboral. Así, se buscó el modelo de mayor consistencia y poder predictivo, como también, se indagó acerca de la influencia que las características (tanto individuales como familiares y laborales) de las personas activas tienen sobre dicha predicción. Se utilizaron tres estimadores bajo un análisis comparativo: dos modelos de clasificación de machine learning (ML), el estimador lasso y el método XGBoost, y la regresión logit que se utiliza comúnmente en la literatura de transiciones laborales.

En cuanto a la comparación de modelos, el estimador XGBoost presenta un poder predictivo significativamente mayor a logit, aunque se resalta la alta competitividad de este último, cuyos valores AUC-ROC son cercanos a los producidos en XGBoost y significativamente mayores a lasso. Resultado que coincide con trabajos previos ([Salganik et al., 2020](#); [Bhattacharyya et al., 2011](#); [Whitrow et al., 2009](#); [Pesantez-Narvaez et al., 2019](#); [Khanna et al., 2015](#); [Jang Su-Kyeong, 2020](#); [Lee et al., 2018](#)). El mejor modelo XGBoost tiene un 85% de probabilidades de que si se eligen dos observaciones al azar, una de “Salida laboral” y otra de “No Salida”, la probabilidad predicha para la primera sea mayor que para la “No salida”, lo que equivale a 2 puntos porcentuales más que logit. Este modelo que mejor desempeño tuvo proyecta una salida del 13% de los individuos activos del 2017, a generarse en el siguiente año, de los cuales 11 puntos porcentuales son mujeres. No obstante, tanto la predicción del mejor modelo, como las demás estimaciones presentan un error de predicción no despreciable, resultado que según [Salganik et al. \(2020\)](#) se atribuye fuertemente a la dificultad de predecir patrones sociales y débilmente a la técnica utilizada.

Dentro de las variables más influyentes se encuentra el género. La inclusión de esta variable aumentó significativamente la *performance* predictiva de los tres modelos (0.811 a 0.834 en logit, 0.803 a 0.827 en lasso y 0.835 a 0.85 en xgboost). Por lo que incluir esta *dummy* es igual o más importante que aumentar la sofisticación del modelo. A su vez, utilizando diferentes técnicas de selección dicho predictor se encuentra entre los más influyentes de la predicción. En el caso de logit la variable de género es la segunda más importante, mientras que en la regresión lasso se ubica en el primer lugar. En XGBoost, de los dos métodos utilizados uno lo posiciona en segundo lugar (método de Ganancia) y el otro en

el primer lugar (importancia SHAP). Dando como resultado que en casi todos los casos el género está por delante o detrás del ingreso de la ocupación principal para los trabajadores, es decir, considerar el género es tan importante como considerar el ingreso laboral de una persona.

Los resultados también coinciden con la literatura en cuanto a la dirección en que se produce la brecha entre varones y mujeres. En los tres modelos, el hecho de ser mujer aumenta la probabilidad predicha de salir del mercado laboral en el siguiente período, en comparación con los varones. En el caso de logit, ser mujer activa aumenta casi 7 puntos porcentuales la probabilidad de salida respecto a su contraparte masculina. A su vez, al interactuar el género con variables como la situación conyugal o la presencia de niños menores de 10 años, la asociación es positiva, es decir, ser mujer casada o vivir en un hogar con al menos un niño menor de 10 años aumenta la probabilidad de salida. Resultado que indica que las condiciones culturales y sociales juegan un papel importante en la generación de la brecha entre varones y mujeres activas. Lo que sugiere la persistencia de roles tradicionales en el hogar, comportamiento que restringe la posibilidad en las mujeres de gozar de una vida laboral plena, como también impide una administración igualitaria entre trabajo y hogar para ambos sexos. Por lo que este trabajo destaca la necesidad de considerar como factor central el género para el estudio de transiciones laborales, debido a que a diferencia de determinantes como el ingreso o educación, la presencia de desigualdades entre mujeres y varones es una condición injusta, y por lo tanto no deseable en pos de lograr un desarrollo sostenible a largo plazo.

Por supuesto, cabe aclarar que dichos resultados deben tomarse con cautela, debido a que los mismos presentan una connotación exploratoria, no así hallazgos de índole causal. En consecuencia, el efecto del género no fue completamente aislado de los demás factores, los cuales posiblemente estén correlacionados.

Además del género, se encuentran otras características relevantes que se alinean con los hallazgos de la literatura de movilidad laboral. La edad, educación, formalidad, antigüedad en el puesto e ingresos se relacionan negativamente con la probabilidad de salir de la fuerza de trabajo. La posibilidad de acumular experiencia, habilidades, educación e ingresos permitiría gozar de una mayor estabilidad y reducir así la chance de interrumpir los vínculos laborales ([Theeuwes et al., 1990](#); [Nilsen et al., 2000](#); [Tasci y Tansel, 2005](#)).

References

- Bergstra, J. y Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305.
- Bhattacharyya, S., Jha, S., Tharakunnel, K., y Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613. On quantitative methods for detection of financial fraud.
- Brefeld, U. y Scheffer, T. (2005). {AUC} maximizing support vector learning.
- CEDLAS y BID (2019). Participación laboral femenina. ¿qué explica las brechas entre países? *Banco Interamericano de Desarrollo*.
- Cerrutti, M. (2000). Determinantes de la participación intermitente de las mujeres en el mercado de trabajo del area metropolitana de buenos aires. *Desarrollo Económico*, 39(156):619–638.
- Chen, T. y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., y Li, Y. (2021). *xgboost: Extreme Gradient Boosting*. R package version 1.4.1.1.
- DeLong, E. R., DeLong, D. M., y Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845.
- Ferreira, F. H. G. y Gignoux, J. (2014). The measurement of educational inequality: Achievement and opportunity. *World Bank Economic Review*, 28(2):210–246.
- Hand, D. J. y Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186.
- Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer Series in Statistics.

- Hersch, J. y Stratton, L. S. (1997). Housework, fixed effects, and wages of married workers. *The Journal of Human Resources*, 32(2):285–307.
- James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Science.
- Jang Su-Kyeong, Chang Jun Young, L. J. S. L. E.-J. K. Y.-H. H. J. H. C. D.-I. C. H. J. C. J.-K. Y. K. H. J. J.-M. A. S. H. K. D.-E. S. S.-I. L. J. H. P. K.-P. K. S. U. K. J. S. K. D.-W. (2020). Reliability and clinical utility of machine learning to predict stroke prognosis: Comparison with logistic regression. *J Stroke*, 22(3):403–406.
- Keith, K. y McWilliams, A. (1997). Job mobility and gender-based wage growth differentials. *Economic Inquiry*, 35(2):320–33.
- Khanna, D., Sahu, R., Baths, V., y Deshpande, B. (2015). Comparative study of classification techniques (svm , logistic regression and neural networks) to predict the prevalence of heart disease.
- Kleinberg, J., Ludwig, J., Mullainathan, S., y Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.
- Lee, H.-C., Yoon, S. B., Yang, S.-M., Kim, W. H., Ryu, H.-G., Jung, C.-W., Suh, K.-S., y Lee, K. H. (2018). Prediction of acute kidney injury after liver transplantation: Machine learning approaches vs. logistic regression model. *Journal of Clinical Medicine*, 7(11).
- Ling, C., Huang, J., y Zhang, H. (2003). Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI*.
- Marchionni, M., Gasparini, L., y Edo, M. (2019). *Brechas de Género en América Latina. Un Estado de Situación*.
- Marston, S. T., Feldstein, M., y Hymans, S. H. (1976). Employment instability and high unemployment rates. *Brookings Papers on Economic Activity*, 1976(1):169–210.
- McCullagh, P. y Nelder, J. (1989). Generalized linear models. *2nd ed. New York: Chapman and Hall*.
- Molnar, C. (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub.

- Mullainathan, S. y Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.
- Nilsen, i. A., Risa, A. E., y Torstensen, A. (2000). Transitions from employment among young norwegian workers. *Journal of Population Economics*, 13(1):21–34.
- Pesantez-Narvaez, J., Guillen, M., y Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7(2).
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A. C., Pentland, A., Polimis, K., Raes, L., Rigobon, D. E., Roberts, C. V., Stanescu, D. M., Suhara, Y., Usmani, A., Wang, E. H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R. B., Argyle, L. P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S. P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I. M., Hummel, L. M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Liu, N., Möser, M., Mack, A. E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K. M., Porter, E., Porter, K. E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J. D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M. K., Woon, W. L., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B. E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B. M., Watts, D. J., y McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403.
- Simon, N., Friedman, J., Hastie, T., y Tibshirani, R. (2011). Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.
- Soules, L. M. (2020). Un modelo de aprendizaje automático orientado a predecir mora

crediticia sobre la base de datos públicos, abiertos y masivos: Desarrollo, evaluación e implicancias prácticas para el mercado crediticio argentino. *Banco Central de la República Argentina*.

Tasci, H. M. y Tansel, A. (2005). Unemployment and transitions in the turkish labor market: Evidence from individual level data. IZA Discussion Papers 1663.

Theeuwes, J., Kerkhofs, M., y Lindeboom, M. (1990). Transition intensities in the dutch labour market 1980–85. *Applied Economics*, 22(8):1043–1061.

Theodossiou, I. (2002). Factors affecting the job-to-joblessness turnover and gender. *LABOUR*, 16(4):729–746.

Theodossiou, I. y Zangelidis, A. (2009). Should i stay or should i go? the effect of gender, education and unemployment on labour market transitions. *Labour Economics*, 16(5):566–577.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Whitrow, C., Hand, D. J., Juszcak, P., Weston, D., y Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1):30–55.

Woltermann, S. (2002). Job-search methods and labor market transitions - some empirical evidence from brazil. *Ibero-America Institute for Economic Research, Ibero America Institute for Econ. Research (IAI) Discussion Papers*.

A Anexo: Predictores

A.1 Características sociodemográficas

- aglomerado: aglomerado donde vive el entrevistado
 - La_Plata (referencia)
 - Bahia_Blanca
 - Rosario
 - Santa_Fe
 - Parana
 - Posadas
 - Resistencia
 - Cdro_Rivadavia
 - Mendoza
 - Corrientes
 - Cordoba
 - Concordia
 - Formosa
 - Neuquen
 - Stgo_del_Estero
 - Jujuy
 - Rio_Gallegos
 - Catamarca
 - Catamarca
 - Salta
 - La_Rioja
 - San_Luis
 - San_Juan
 - Tucuman

- Santa_Rosa
 - Ushuaia
 - CABA
 - GBA
 - Mar_del_Plata
 - Rio_Cuarto
 - San_Nicolas
 - Rawson
 - Viedma
- casado: *dummy* de si la persona está unida o casada.
 - edad: edad de la persona.
 - edad_2: edad al cuadrado.
 - extranjero: *dummy* de si la persona es extranjera.
 - jefe_pareja: relación de parentesco con el/la jefe/a de hogar
 - Jefe_hogar (referencia): categoría de jefatura de hogar.
 - Conyuge_pareja: categoría de cónyuge de hogar.
 - Otros: demás categorías. Incluyen hijo/a, hijastro/a, nieto/a, hermano/a, entre otros.
 - log_ingreso_nolaboral: logaritmo del ingreso no laboral, deflactado según IPC con base en el cuatro trimestre de 2019.
 - log_ingreso_total_individual: logaritmo del ingreso total de cada encuestado, deflactado según IPC con base en el cuatro trimestre de 2019.
 - mujer: representa la *dummy* de si la persona es mujer.
 - region⁵¹

⁵¹ Se optó por utilizar aglomerados para los modelos, que representaría una versión más desagregada del aspecto geográfico. Aunque exploratoriamente se utiliza la región para ver la distribución de la tasa de transición de varones y mujeres.

- GBA_CABA (referencia): Gran Buenos Aires y Ciudad Autónoma de Buenos Aires.
- NOA: Noroeste argentino.
- NEA: Noreste argentino.
- Cuyo: Región centro oeste de Argentina.
- Pampeana: Región centro de Argentina.
- Patagonia: Región sur de Argentina.

A.2 Características laborales

- **antig**: antigüedad en el puesto
 - 1_3_meses (referencia): entre 1 y 3 meses
 - 3_6_meses: entre 3 y 6 meses
 - 6_12_meses: entre 6 y 12 meses
 - 1_5_anios: entre 1 y 5 años
 - mas_5_anios: más de 5 años
 - Otros: desocupados y ocupados que no saben la antigüedad en el puesto
- **calific**: calificación ocupacional
 - Profesional (referencia): calificación profesional
 - Tecnico: nivel técnico
 - Operativo: nivel operativo
 - No_calificado: personal no calificado
 - Otros: categoría que incluye trabajadores no clasificados dentro de alguna calificación o desocupados
- **cat_ocup**
 - Obrero_empleado (referencia): categoría ocupacional obrero o empleado.
 - Cuenta_propia_prof: cuenta propista con nivel educativo alto.
 - Cuenta_propia_noprof: cuenta propista con nivel educativo bajo o medio.

- Patron: categoría ocupacional patrón.
 - Fliar_s_remuneracion: categoría trabajador familiar sin remuneración.
 - Otros: representa a los activos que se encuentran desempleados.
- educ
 - Nivel_bajo (referencia): hasta secundaria incompleta.
 - Nivel_medio: hasta terciaria incompleta.
 - Nivel_alto: terciaria completa.
- finalizacion: trabajo con o sin finalización
 - Tiene_finalizacion (referencia)
 - No_tiene_finaliacion
 - Otros
- formal
 - Formal (referencia): personas con trabajo que se encuentren dentro del sistema de seguridad social.
 - Informal: personas con trabajo fuera del sistema de seguridad social.
 - Otros: desempleados.
- intens_horaria: intensidad horaria.
 - Ocup_plena (referencia): carga horaria del trabajo entre 35 y 45 horas semanales.
 - Subocup_voluntaria: carga horaria menor de 35 horas semanales sin intención de trabajar más horas.
 - Subocup_involuntaria: carga horaria menor de 35 horas semanales con intención de trabajar más horas.
 - Sobre_ocup: carga horaria mayor a 45 horas semanales
 - Otros: Desocupados
- jerar: jerarquía ocupacional
 - Direccion (referencia): puesto de dirección

- Cta_propia: cuenta propista
- Jefe: puesto de jefatura
- Trabaj_asalariado: puesto de empleado o trabajador asalariado
- Otros: categoría que incluye trabajadores no clasificados dentro de alguna jerarquía o desocupados
- log_ingreso_otras_ocupaciones: logaritmo del ingreso de ocupaciones secundarias del trabajador, deflactado según IPC con base en el cuatro trimestre de 2019.
- log_ingreso_real: logaritmo del ingreso de la ocupación principal, deflactado según IPC con base en el cuatro trimestre de 2019.
- ocup_desocup
 - Desocup (referencia): desocupados sin ocupación anterior.
 - Desocup_ocup_anterior: desocupados con ocupación anterior.
 - Ocupado: ocupados.
- sector_2: sector de la ocupación principal del ocupado y desocupado con experiencia laboral agregado según CAES Mercosur 1.0 versión Argentina. ⁵²
 - S1 (referencia): Agricultura, Ganadería, Caza, Silvicultura y Pesca
 - S2: Explotación de Minas y Canteras
 - S3: Industria Manufacturera
 - S4: Suministro de Electricidad, Gas, Vapor y Aire Acondicionado
 - S5: Suministro De Agua; Alcantarillado, Gestión de Desechos y Actividades de Saneamiento
 - S6: Construcción
 - S7: Comercio al por Mayor y al por Menor; Reparación de Vehículos Automotores y Motocicletas
 - S8: Transporte y Almacenamiento
 - S9: Alojamiento y Servicios de Comidas
 - S10: Información y Comunicación

⁵² Disponible en [INDEC](#).

- S11: Actividades Financieras y de Seguros
 - S12: Actividades Inmobiliarias
 - S13: Actividades Profesionales, Científicas y Técnicas
 - S14: Actividades Administrativas y Servicios de Apoyo
 - S15: Administración Pública y Defensa; Planes de Seguro Social Obligatorio
 - S16: Enseñanza
 - S17: Salud Humana y Servicios Sociales
 - S18: Artes, Entretenimiento y Recreación
 - S19: Otras Actividades de Servicios
 - S20: Actividades de los Hogares como Empleadores de Personal Doméstico;
Actividades de los Hogares como Productores de Bienes o Servicios para Uso Propio
 - S21: Otros
 - S22: Ninguno
- **tam_establecimiento:** cantidad de personas que trabajan en el mismo establecimiento
 - **Hasta_5** (referencia): menos de 5 personas
 - **Entre_6_40:** entre 6 y 40 personas
 - **Mas_40:** más de 40 personas
 - **Otros:** corresponde de ocupados que no sabe y al grupo de desocupados
 - **tecno_ocup:** tecnología ocupacional
 - **Sin_maquina** (referencia): sin operación de máquina
 - **Con_maquina:** con operación de máquina y equipos electromecánicos
 - **Equipos_informaticos:** operación de sistemas y equipos informatizados
 - **Otros:** categoría que incluye trabajadores no clasificados dentro de alguna tecnología o desocupados
 - **tiempo_busqueda:** tiempo de búsqueda laboral de los desocupados
 - **Menos_1_mes** (referencia)

- 1_3_meses
 - 3_6_meses
 - 6_12_meses
 - Mas_1_anio
 - Otros
- tiempo_ocup_anterior: período de tiempo en el trabajo anterior de los desocupados con ocupación anterior
 - Menos_1_mes (referencia)
 - 1_3_meses
 - 3_6_meses
 - 6_12_meses
 - 1_3_anios
 - Mas_3_anios
 - Otros
 - tipo: razón social del establecimiento donde trabaja el ocupado
 - Sociedad_jurídica (referencia)
 - Actividad_familiar
 - Otro_tipo
 - Otros
 - total_horas_op: total de horas trabajadas en la ocupación principal⁵³.
 - total_horas_os: total de horas trabajadas en otras ocupaciones. ⁵⁴

A.3 Características del hogar

- calidad_vivienda: calidad de los materiales de la vivienda.⁵⁵
 - Suficiente (referencia)

⁵³ A los desocupados se les asigna un valor de cero horas trabajadas.

⁵⁴ A los desocupados se les asigna un valor de cero horas trabajadas.

⁵⁵ Más detalles de la construcción del indicador puede encontrarse en [INDEC](#).

- `Parcialmente_insuficiente`
 - `Insuficiente`
 - `Otros`
- `deccfr`: decil del ingreso per cápita familiar del hogar.
 - `fuentegto_ahorro`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de gastar lo que tenían ahorrado.
 - `fuentegto_alquiler`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de algún alquiler (por una vivienda, terreno, oficina, etc.) de su propiedad.
 - `fuentegto_banco`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de pedir préstamos a bancos, financieras, etc.
 - `fuentegto_fliares`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de pedir préstamos a familiares/amigos.
 - `fuentegto_indemn_despido`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de alguna indemnización por despido.
 - `fuentegto_intereses_pf`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de intereses o rentas por plazos fijos/inversiones.
 - `fuentegto_jub_pens`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de alguna jubilación o pensión.
 - `fuentegto_otro_negocio`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de ganancias de algún negocio en el que no trabajan.
 - `fuentegto_seguro_desemp`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de algún seguro de desempleo.
 - `fuentegto_subsidio`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de algún subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.
 - `fuentegto_tarjcredito`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de comprar en cuotas o al fiado con tarjeta de crédito o libreta.

- `fuentegto_trabajo`: dummy con 1 si en los últimos tres meses las personas del hogar han vivido de lo que ganan en el trabajo.
- `log_ipcf_real`: logaritmo del ingreso ingreso per cápita familiar, deflactado según IPC con base en el cuatro trimestre de 2019.
- `menor1_tot_hogar`: cantidad de personas recién nacidas.
- `menor6_tot_hogar`: cantidad de personas menores de 6 años.
- `menor10_tot_hogar`: cantidad de personas menores de 10 años en el hogar.
- `miembros`: cantidad de miembros en el hogar.⁵⁶
- `tipo_vivienda`: tipo de vivienda del hogar
 - Casa(referencia)
 - Departamento
 - Pieza_inquilinato
 - Pieza_hotel_pension
 - Local_noconstruido_p_habitacion
 - Otros

A.4 Características locales

- `tasa_desoc`: tasa de desocupación en los aglomerados. Entendiéndose desocupación como el ratio entre la población desocupada sobre la población económicamente activa de entre 25 y 59 años de edad.
- `tasa_empleo`: tasa de empleo por aglomerado. Entendiéndose esta tasa como el ratio entre la población ocupada sobre la población total de entre 25 y 59 años.
- `tasa_inactividad`: tasa de inactividad por aglomerado. Entendiéndose inactividad como el ratio entre la población que no trabaja ni busca trabajo sobre la población total de entre 25 y 59 años.
- `villa`: dummy de si la vivienda se ubica en una villa de emergencia.

⁵⁶ Cabe aclarar que debido a las pocas observaciones que se encuentran en hogares con más de diez miembros, se opta por asignar el valor 10 a todos los hogares que presentan más de 10 personas en el hogar. Por lo que dicho valor representa hogares con al menos 10 miembros.

A.5 Características coyunturales

- **emae**: estimador mensual de la actividad económica elaborado por el Instituto de Estadísticas y Censos de la República Argentina.
- **lider**: índice líder elaborado por la Universidad Torcuato Di Tella.
- **tcnm**: tipo de cambio real nominal elaborado por el Banco Central de la República Argentina.

B Anexo: Modelos

B.1 Regresión logit: proceso de optimización

Sea la *log-likelihood* para N observaciones y $\theta = \beta_0 + \sum_{p=1}^p X_{ip}\beta_p$

$$\ell(\theta) = \sum_{i=1}^N \log \pi_{y_i}(x_i; \theta), \quad (16)$$

Para una variable binaria y_i toma los valores 0 y 1.

Luego, podemos establecer la función *log-likelihood* de la siguiente forma

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^N \{y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)\} \\ &= \sum_{i=1}^N \{y_i \log \pi_i + \log(1 - \pi_i) - y_i \log(1 - \pi_i)\} \\ &= \sum_{i=1}^N \left\{ \log(1 - \pi_i) + y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) \right\} \\ &= \sum_{i=1}^N \left\{ \log \left(\frac{1}{1 + \exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p)} \right) + y_i \left(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p \right) \right\} \\ &= \sum_{i=1}^N \left\{ y_i \left(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p \right) - \log(1 + \exp(\beta_0 + \sum_{p=1}^p X_{ip}\beta_p)) \right\} \end{aligned} \quad (17)$$

Así, se llega a la función objetivo a maximizar.

B.2 XGBoost: término de penalización

Se define el término de penalización como

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \frac{2}{2} \quad (18)$$

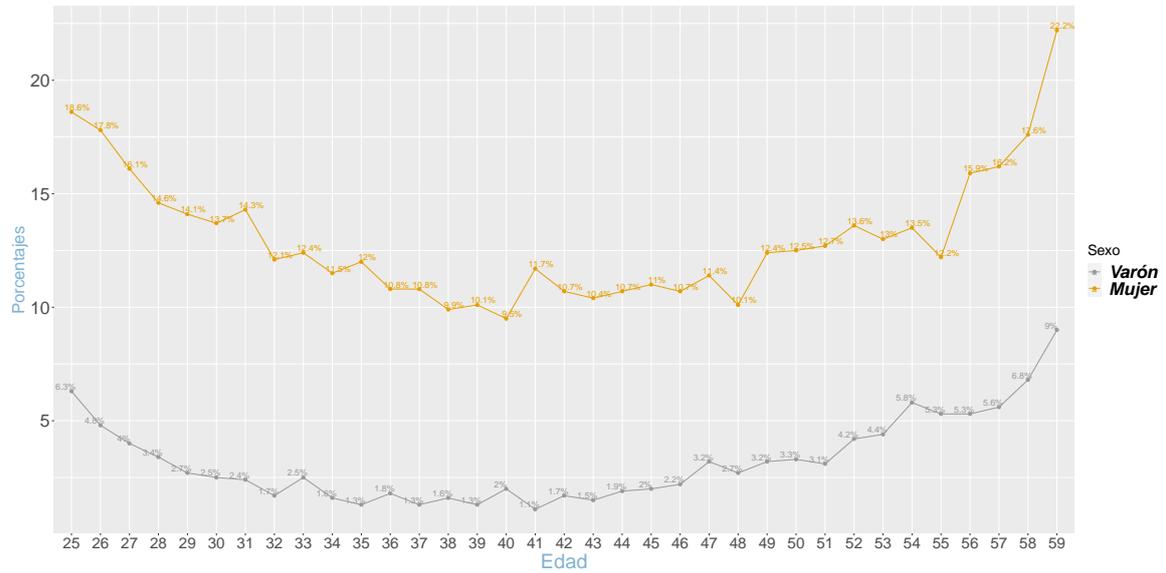
donde el hiperparámetro γ representa el *pruning*, el cual regula la profundidad del árbol y, λ es el hiperparámetro de regularización.⁵⁷ El hiperparámetro *pruning* es una técnica de ML que reduce el tamaño del árbol de decisión, removiendo nodos cuya característica correspondiente tiene menos influencia sobre la predicción final de la variable dependiente. Este último procedimiento reduce la complejidad del modelo y, así, corrige el sobreajuste.

⁵⁷ Si bien existen muchas formas de definir la penalización, la Ecuación utilizada se debe al buen funcionamiento en la práctica. Más información disponible en [XGBoost](#).

C Anexo: Gráficos y tablas

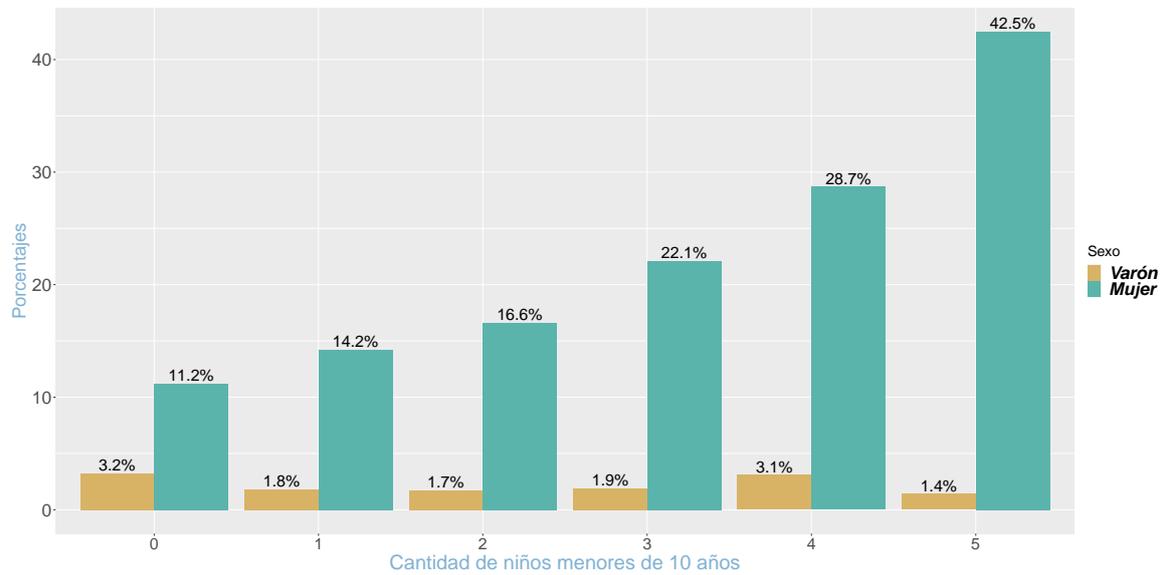
C.1 Descriptivos

Figura 21: Proporción de transiciones a la inactividad de mujeres y varones según edad



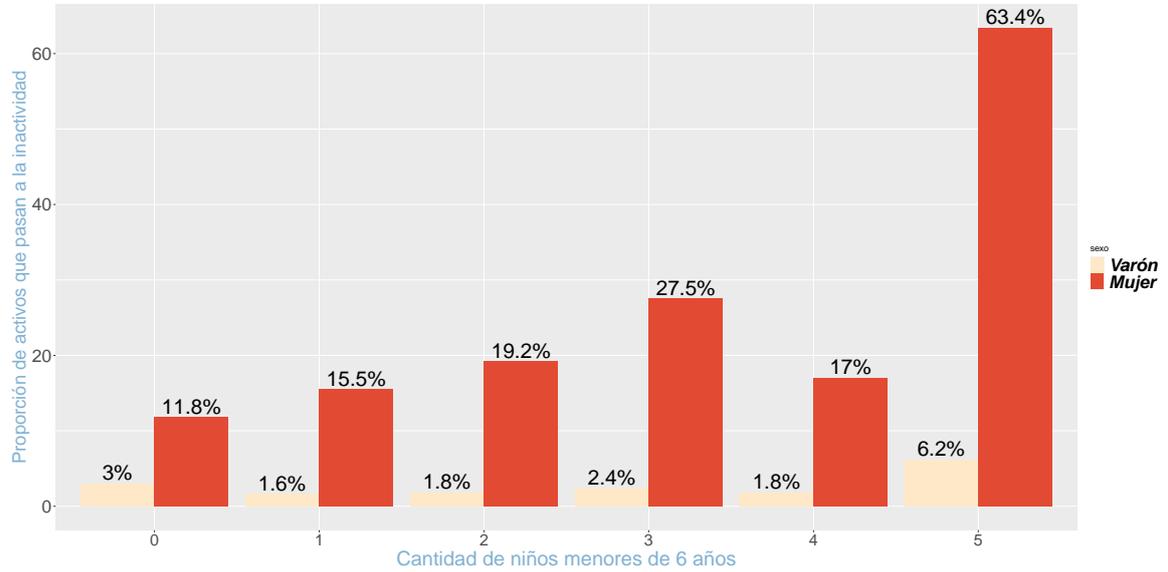
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 22: Proporción de transiciones a la inactividad de mujeres y varones según presencia de menores de 10 años en el hogar



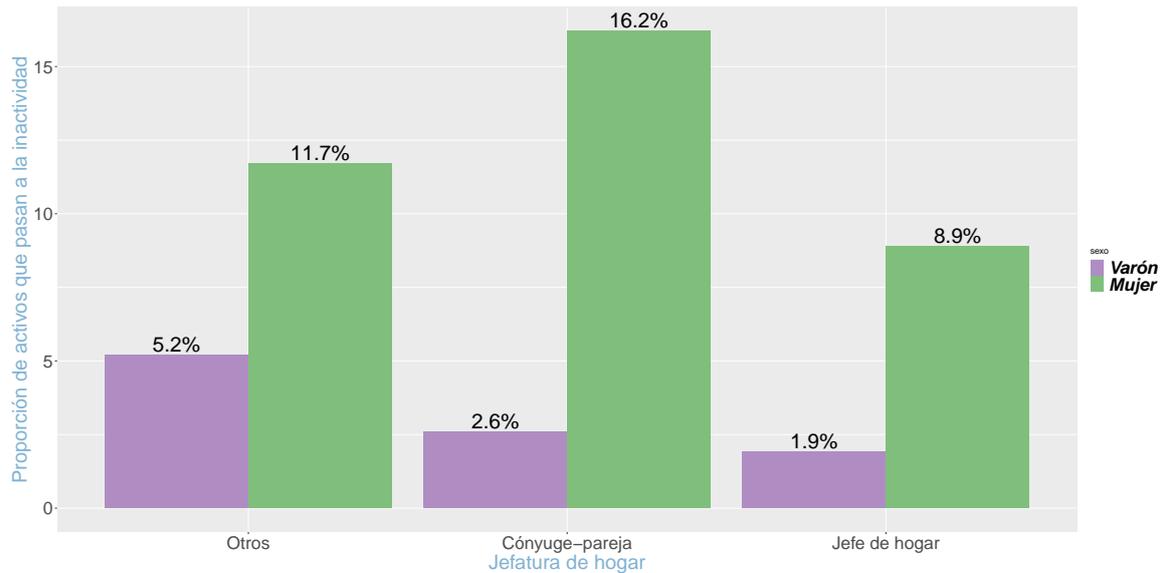
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 23: Proporción de transiciones a la inactividad de mujeres y varones según presencia de menores de 6 años en el hogar



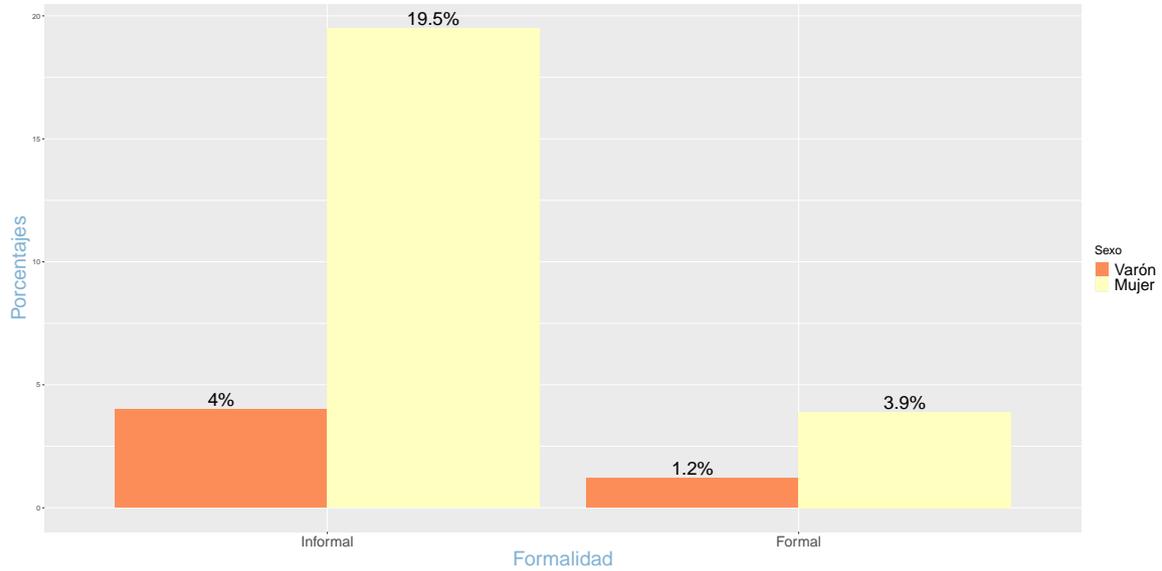
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 24: Proporción de transiciones a la inactividad de mujeres y varones según jefatura de hogar



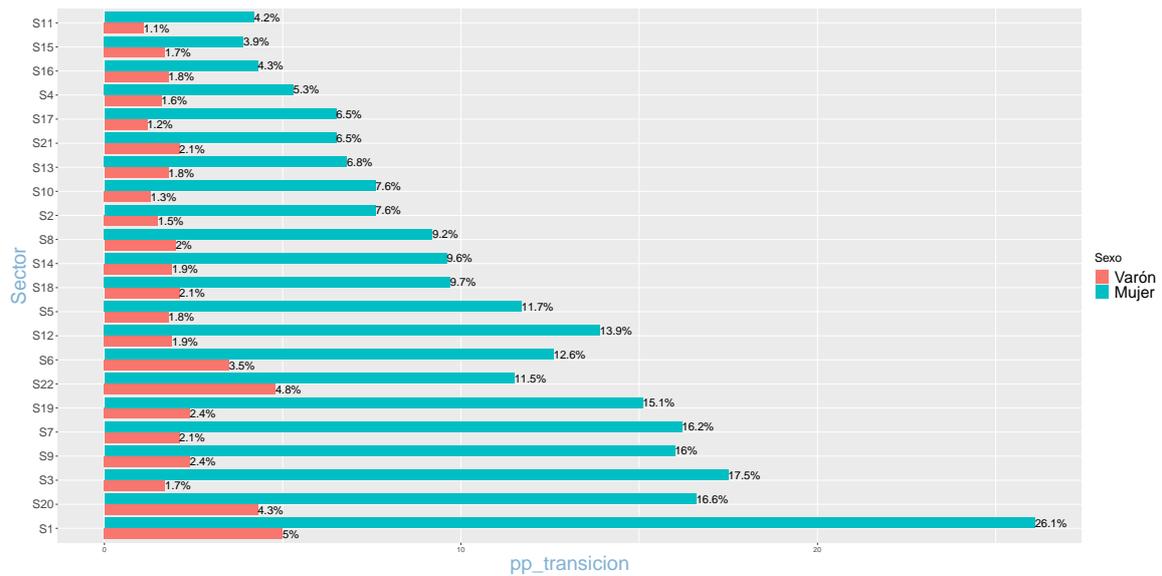
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 25: Proporción de transiciones a la inactividad de mujeres y varones según formalidad



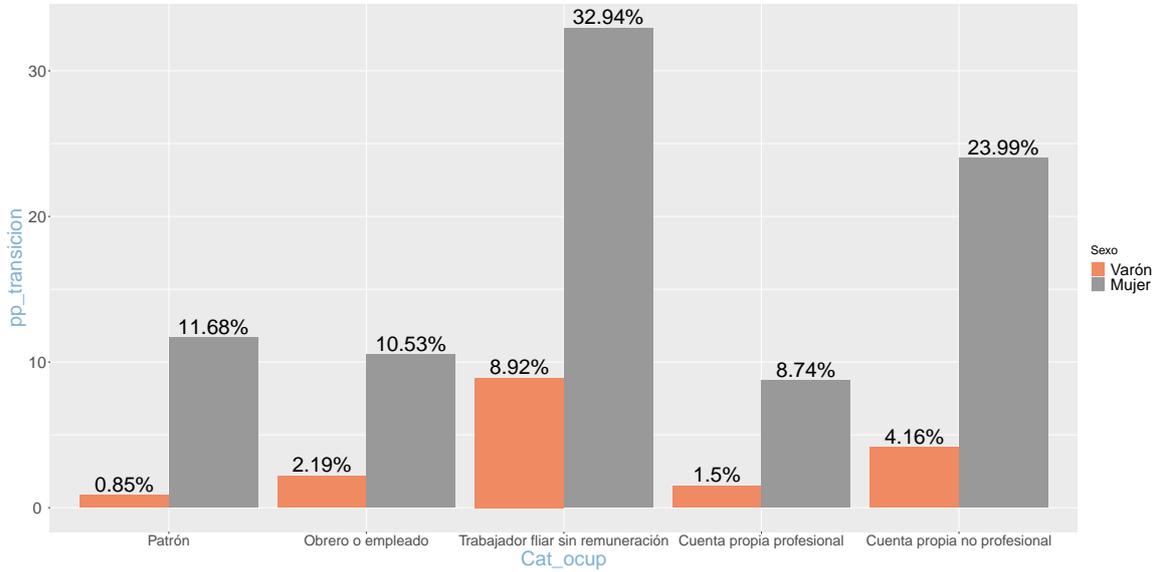
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 26: Proporción de transiciones a la inactividad de mujeres y varones según sector productivo



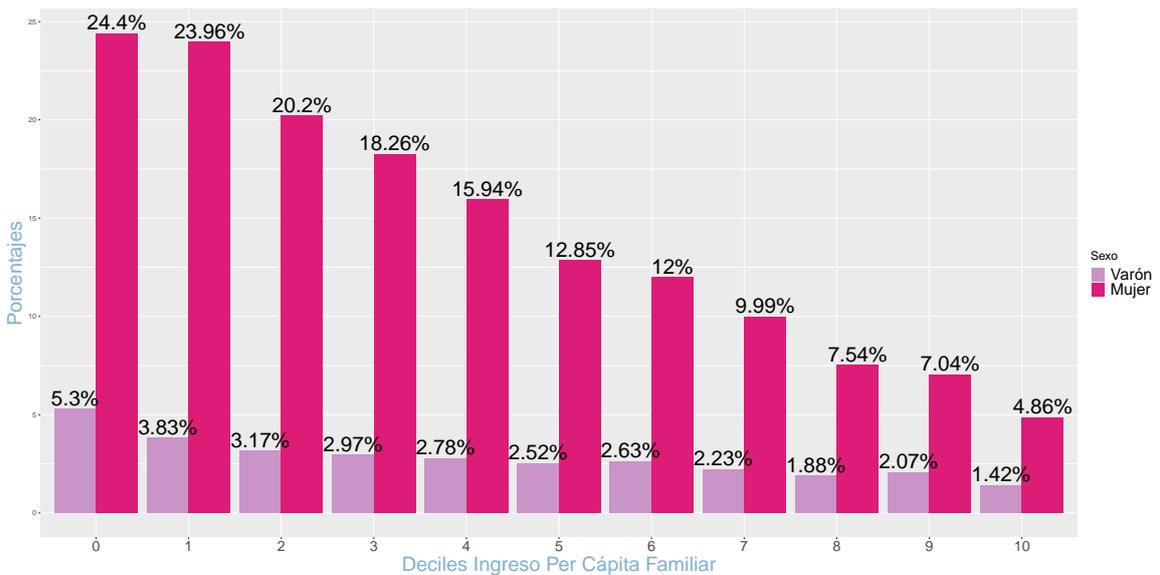
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 27: Proporción de transiciones a la inactividad de mujeres y varones según categoría ocupacional



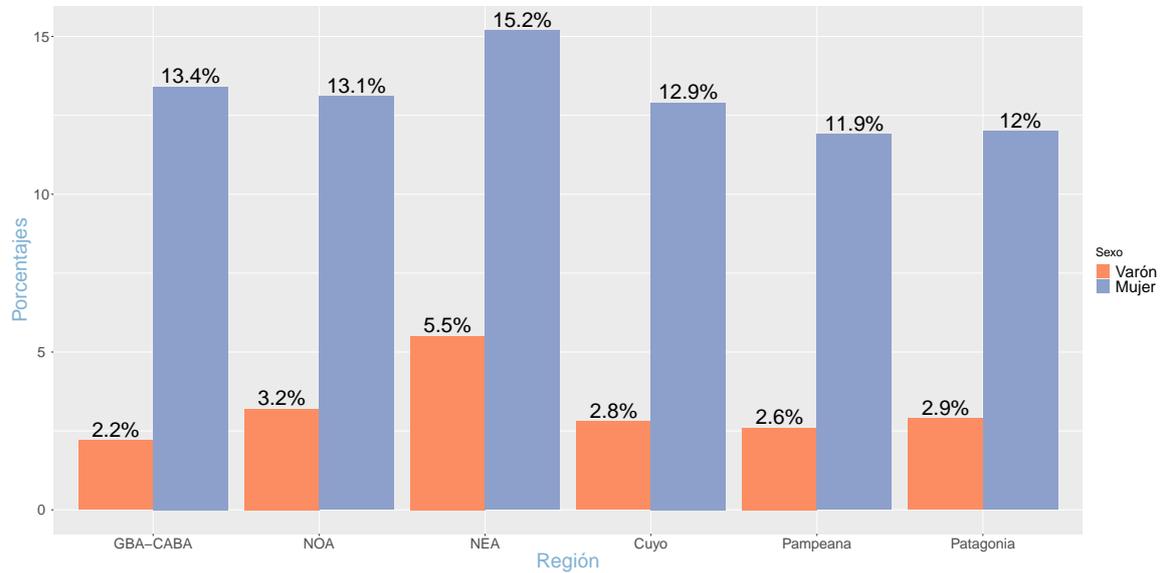
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 28: Proporción de transiciones a la inactividad de mujeres y varones según decil del Ingreso Per Cápita Familiar



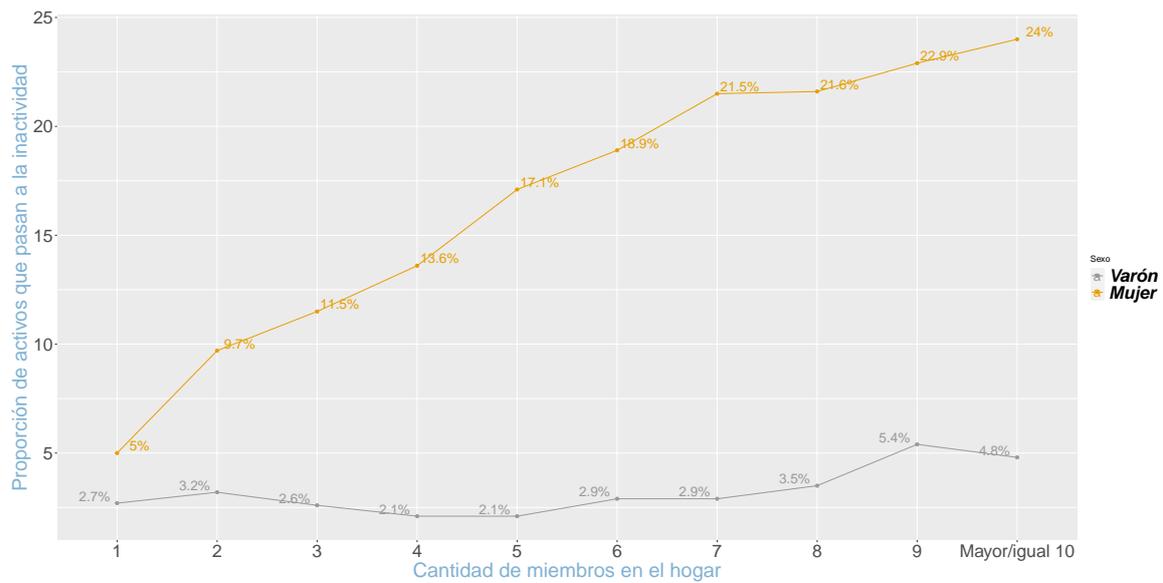
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 29: Proporción de transiciones a la inactividad de mujeres y varones según región



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

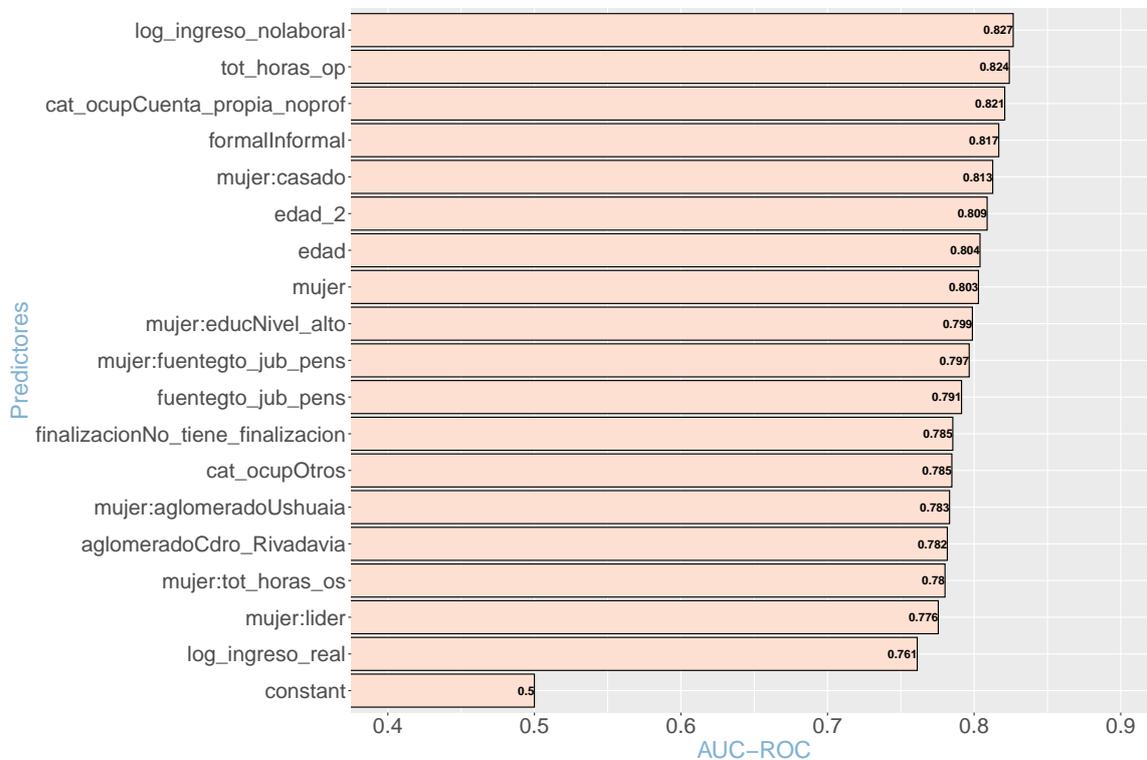
Figura 30: Proporción de transiciones a la inactividad de mujeres y varones según cantidad de miembros en el hogar



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

C.2 Regresión logística

Figura 31: Método de selección *forward stepwise selection* para el modelo con interacciones entre el género y las demás predictoras



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2017

Tabla 4: Regresión logit. Modelo con género sin interacciones

	log-odds
(Intercept)	2.0373 (1.7165)
sector_2S2	-0.3092 (0.2220)
sector_2S3	-0.1795* (0.0941)
sector_2S4	0.2304 (0.2051)
sector_2S5	-0.3254* (0.1976)
sector_2S6	-0.4346*** (0.0999)

	log-odds
sector_2S7	-0.1964** (0.0910)
sector_2S8	-0.3218*** (0.1149)
sector_2S9	-0.1751* (0.1016)
sector_2S10	-0.1650 (0.1397)
sector_2S11	-0.3849*** (0.1449)
sector_2S12	0.0891 (0.1781)
sector_2S13	-0.3144*** (0.1043)
sector_2S14	-0.1710 (0.1301)
sector_2S15	-0.6990*** (0.1013)
sector_2S16	-0.6663*** (0.1013)
sector_2S17	-0.5013*** (0.1000)
sector_2S18	-0.2186* (0.1149)
sector_2S19	-0.4183*** (0.1008)
sector_2S20	0.0772 (0.1095)
sector_2S21	-0.3091 (0.2809)
sector_2S22	-0.2851 (0.3813)

	log-odds
miembros	0.0589*** (0.0080)
menor10_tot_hogar	0.0248 (0.0208)
menor6_tot_hogar	0.0078 (0.0273)
menor1_tot_hogar	-0.1030** (0.0487)
mujer	1.1461*** (0.0328)
jefe_parejaOtros	0.3013*** (0.0402)
jefe_parejaConyuge_pareja	0.6577*** (0.0334)
edad_2	0.0028*** (0.0001)
edad	-0.2155*** (0.0099)
casado	0.1870*** (0.0341)
educNivel_medio	-0.1154*** (0.0255)
educNivel_alto	-0.5944*** (0.0471)
cat_ocupOtros	2.1214 (1.8302)
cat_ocupPatron	-0.3616** (0.1603)
cat_ocupCuenta_propia_noprof	0.4234*** (0.0732)
cat_ocupFliar_s_remuneracion	0.1317 (0.1481)

	log-odds
cat_ocupCuenta_propia_prof	0.2482** (0.1039)
formalOtros	0.0580 (0.1233)
formalInformal	0.8274*** (0.0408)
ocup_desocupDesocup_ocup_anterior	-0.0445 (0.0640)
ocup_desocupOcupado	0.3250 (2.0446)
extranjero	-0.0220 (0.0572)
tot_horas_os	-0.0253*** (0.0045)
tot_horas_op	-0.0196*** (0.0012)
intens_horariaOtros	0.2981 (0.6014)
intens_horariaSubocup_voluntaria	0.1953*** (0.0427)
intens_horariaSubocup_involuntaria	0.0715 (0.0506)
intens_horariaSobre_ocup	0.2312*** (0.0419)
antigOtros	-0.1551 (0.2243)
antig3_6_meses	0.1613** (0.0657)
antig6_12_meses	0.1067* (0.0610)
antig1_5_anios	-0.1550*** (0.0467)

	log-odds
antigmas_5_anios	-0.4319*** (0.0497)
jerarqDireccion	0.8214 (0.8104)
jerarqJefe	0.1656 (0.8175)
jerarqTrabaj_asalariado	-0.1811 (0.1167)
jerarqOtros	0.5723 (1.4255)
tecno_ocupCon_maquina	-0.0319 (0.0526)
tecno_ocupEquipos_informaticos	-0.1167** (0.0473)
tecno_ocupOtros	-0.3080 (0.8016)
calificTecnico	0.2489*** (0.0715)
calificOperativo	0.3490*** (0.0736)
calificNo_calificado	0.3921*** (0.0797)
calificOtros	-0.0872 (0.6866)
tam_establecimientoEntre_6_40	-0.0201 (0.0416)
tam_establecimientoMas_40	-0.1380*** (0.0498)
tam_establecimientoOtros	0.1158* (0.0635)
tipo_viviendaDepartamento	-0.0421 (0.0330)

	log-odds
tipo_viviendaPieza_inquilinato	-0.0803 (0.1663)
tipo_viviendaPieza_hotel_pension	-0.1334 (0.4029)
tipo_viviendaLocal_noconstruido_p_habitacion	-1.3476* (0.7438)
tipo_viviendaOtros	-0.1022 (0.2218)
lugar_trabajoPuesto_kiosco	0.3257* (0.1677)
lugar_trabajoVehiculo	0.0891 (0.1769)
lugar_trabajoVehiculo_transporte_personas	0.1687* (0.0949)
lugar_trabajoObra_construccion	0.1002 (0.0983)
lugar_trabajoVivienda_propia	0.5294*** (0.0490)
lugar_trabajoVivienda_socio_patron	0.3893** (0.1912)
lugar_trabajoDomicilio_clientes	0.1731*** (0.0478)
lugar_trabajoCalle	0.2175*** (0.0617)
lugar_trabajoOtro_lugar	0.4216*** (0.1204)
lugar_trabajoOtros	-0.7498*** (0.2393)
tipoOtros	0.3934 (0.2511)
tipoActividad_familiar	0.4955* (0.2539)

	log-odds
tipoOtro_tipo	0.0190 (0.2825)
finalizacionNo_tiene_finalizacion	-0.1802*** (0.0468)
finalizacionOtros	0.0751 (0.0549)
tiempo_búsqueda1_3_meses	-0.0061 (0.1051)
tiempo_búsqueda3_6_meses	-0.0068 (0.1213)
tiempo_búsqueda6_12_meses	0.0681 (0.1217)
tiempo_búsquedaMas_1_año	0.0692 (0.1041)
tiempo_búsquedaOtros	0.7186 (0.8549)
tiempo_ocup_anterior1_3_meses	0.0226 (0.0985)
tiempo_ocup_anterior3_6_meses	0.1593 (0.1213)
tiempo_ocup_anterior6_12_meses	0.3346*** (0.1283)
tiempo_ocup_anterior1_3_años	0.5103*** (0.1189)
tiempo_ocup_anteriorMas_3_años	-1.2270 (1.8574)
tiempo_ocup_anteriorOtros	-1.4819 (1.8548)
aglomeradoBahia_Blanca	-0.0258 (0.1009)
aglomeradoRosario	0.0883 (0.0868)

	log-odds
aglomeradoSanta_Fe	0.1577* (0.0935)
aglomeradoParana	0.1268 (0.0967)
aglomeradoPosadas	0.4960*** (0.0937)
aglomeradoResistencia	1.1873*** (0.0872)
aglomeradoCdro_Rivadavia	0.4944*** (0.0995)
aglomeradoMendoza	0.2952*** (0.0874)
aglomeradoCorrientes	0.2796*** (0.0908)
aglomeradoCordoba	0.1639* (0.0860)
aglomeradoConcordia	0.5662*** (0.0908)
aglomeradoFormosa	0.9455*** (0.0947)
aglomeradoNeuquen	0.4873*** (0.0976)
aglomeradoStgo_del_Estero	0.4527*** (0.0929)
aglomeradoJujuy	0.2183** (0.0917)
aglomeradoRio_Gallegos	0.2324** (0.1081)
aglomeradoCatamarca	0.2681*** (0.0917)
aglomeradoSalta	0.0133 (0.0879)

	log-odds
aglomeradoLa_Rioja	0.0236 (0.0976)
aglomeradoSan_Luis	0.2309** (0.0970)
aglomeradoSan_Juan	0.2639*** (0.0927)
aglomeradoTucuman	0.1080 (0.0851)
aglomeradoSanta_Rosa	-0.4175*** (0.1228)
aglomeradoUshuaia	0.4921*** (0.0997)
aglomeradoCABA	-0.0100 (0.0947)
aglomeradoGBA	0.1141 (0.0754)
aglomeradoMar_del_Plata	-0.1194 (0.0966)
aglomeradoRio_Cuarto	-0.0965 (0.1026)
aglomeradoSan_Nicolas	0.3265*** (0.0997)
aglomeradoRawson	0.1158 (0.1041)
aglomeradoViedma	0.1135 (0.1044)
tasa_desoc	-0.6541 (0.4503)
tasa_empleo	-2.5773 (1.6005)
tasa_inactividad	-2.3031 (1.5362)

	log-odds
log_ingreso_real	-0.1691*** (0.0177)
log_ipcf_real	0.2069** (0.0961)
log_ingreso_otras_ocupaciones	-0.1149*** (0.0145)
log_ingreso_nolaboral	0.1468*** (0.0094)
log_ingreso_total_individual	-0.0502*** (0.0157)
lider	-0.0019 (0.0016)
emae_orig	0.0062*** (0.0015)
tcnm	-0.0086 (0.0364)
villa	0.0937 (0.0820)
calidad_viviendaParcialmente_insuficiente	0.0553 (0.0589)
calidad_viviendaInsuficiente	0.0572 (0.0621)
calidad_viviendaOtros	0.1414 (0.0970)
fuentegto_trabajo	-0.1087* (0.0645)
fuentegto_jub_pens	0.2229*** (0.0280)
fuentegto_indemn_despido	-0.2961** (0.1299)
fuentegto_seguro_desemp	-0.3473*** (0.1185)

	log-odds
fuentegto_subsidio	-0.2580*** (0.0350)
fuentegto_alquiler	-0.0473 (0.0674)
fuentegto_otro_negocio	-0.1092 (0.2515)
fuentegto_intereses_pf	-0.2267 (0.2461)
fuentegto_ahorro	-0.0302 (0.0277)
fuentegto_fliares	-0.0053 (0.0341)
fuentegto_banco	-0.0170 (0.0361)
fuentegto_tarjcredito	0.0259 (0.0228)
deccfr0	0.6438* (0.3513)
deccfr2	0.0848* (0.0476)
deccfr3	0.0839 (0.0595)
deccfr4	0.1159* (0.0693)
deccfr5	0.0856 (0.0786)
deccfr6	0.1190 (0.0855)
deccfr7	0.1708* (0.0941)
deccfr8	0.1520 (0.1043)

	log-odds
deccfr9	0.2694** (0.1160)
deccfr10	0.2449* (0.1398)
AIC	67837.0704
BIC	69520.6330
Log Likelihood	-33750.5352
Deviance	67501.0704
Num. obs.	166243

Fuente: Elaboración propia con base en la EPH 2004-2016. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Tabla 5: Regresión logit. Especificación con las interacciones de género

	log-odds
(Intercept)	-2.6117 (3.7080)
mujer	7.3314* (4.2241)
sector_2S2	-0.4287 (0.2767)
sector_2S3	-0.4300*** (0.1323)
sector_2S4	0.1839 (0.2495)
sector_2S5	-0.4441 (0.2754)
sector_2S6	-0.2924** (0.1279)
sector_2S7	-0.4640*** (0.1266)
sector_2S8	-0.4477*** (0.1488)
sector_2S9	-0.2637

	log-odds
	(0.1613)
sector_2S10	-0.3415
	(0.2163)
sector_2S11	-0.5327**
	(0.2448)
sector_2S12	0.0346
	(0.3278)
sector_2S13	-0.1627
	(0.1471)
sector_2S14	-0.3976**
	(0.1907)
sector_2S15	-0.5421***
	(0.1448)
sector_2S16	-0.5664***
	(0.1752)
sector_2S17	-0.8097***
	(0.2109)
sector_2S18	-0.2836
	(0.1748)
sector_2S19	-0.3430**
	(0.1570)
sector_2S20	0.1030
	(0.2886)
sector_2S21	0.0482
	(0.4224)
sector_2S22	-0.0458
	(0.6238)
miembros	0.0658***
	(0.0147)
menor10_tot_hogar	-0.0919*
	(0.0475)
menor6_tot_hogar	0.0077

	log-odds
	(0.0639)
menor1_tot_hogar	-0.1335
	(0.1095)
jefe_parejaOtros	0.6015***
	(0.0746)
jefe_parejaConyuge_pareja	0.4018***
	(0.0829)
edad_2	0.0032***
	(0.0002)
edad	-0.2292***
	(0.0185)
casado	-0.3403***
	(0.0604)
educNivel_medio	-0.0171
	(0.0489)
educNivel_alto	-0.4771***
	(0.1036)
cat_ocupOtros	1.6430
	(2.6903)
cat_ocupPatron	-0.5276*
	(0.2749)
cat_ocupCuenta_propia_noprof	0.3079***
	(0.1127)
cat_ocupFliar_s_remuneracion	0.4622
	(0.3092)
cat_ocupCuenta_propia_prof	0.1217
	(0.2111)
formalOtros	0.1747
	(0.2137)
formalInformal	0.5109***
	(0.0797)
ocup_desocupDesocup_ocup_anterior	-0.0180

	log-odds
	(0.1139)
ocup_desocupOcupado	-0.0707 (2.9341)
extranjero	-0.0736 (0.1213)
tot_horas_os	-0.0206** (0.0092)
tot_horas_op	-0.0170*** (0.0024)
intens_horariaOtros	0.4187 (1.2188)
intens_horariaSubocup_voluntaria	0.3837*** (0.0837)
intens_horariaSubocup_involuntaria	0.1844* (0.0973)
intens_horariaSobre_ocup	0.2287*** (0.0715)
antigOtros	-0.3614 (0.4135)
antig3_6_meses	0.3586*** (0.1146)
antig6_12_meses	0.1501 (0.1145)
antig1_5_anios	-0.1181 (0.0866)
antigmas_5_anios	-0.2803*** (0.0891)
jerarqDireccion	0.9312 (1.0950)
jerarqJefe	0.4912 (1.0992)
jerarqTrabaj_asalariado	0.0829

	log-odds
	(0.1879)
jerarqOtros	-0.2683 (1.9893)
tecno_ocupCon_maquina	-0.1493 (0.0975)
tecno_ocupEquipos_informaticos	-0.0353 (0.0894)
tecno_ocupOtros	-0.5047 (1.0731)
calificTecnico	0.1470 (0.1353)
calificOperativo	0.2708** (0.1379)
calificNo_calificado	0.2906* (0.1506)
calificOtros	0.8597 (0.9526)
tam_establecimientoEntre_6_40	0.0631 (0.0746)
tam_establecimientoMas_40	-0.0917 (0.0903)
tam_establecimientoOtros	0.1060 (0.1053)
tipo_viviendaDepartamento	-0.0500 (0.0692)
tipo_viviendaPieza_inquilinato	-0.0349 (0.3254)
tipo_viviendaPieza_hotel_pension	-0.3557 (0.7691)
tipo_viviendaLocal_noconstruido_p_habitacion	-8.7059 (44.5436)
tipo_viviendaOtros	0.1285

	log-odds
	(0.3922)
lugar_trabajoPuesto_kiosco	0.6472*
	(0.3425)
lugar_trabajoVehiculo	0.1985
	(0.1980)
lugar_trabajoVehiculo_transporte_personas	0.3015**
	(0.1213)
lugar_trabajoObra_construccion	-0.0175
	(0.1131)
lugar_trabajoVivienda_propia	0.3833***
	(0.1198)
lugar_trabajoVivienda_socio_patron	0.2967
	(0.3217)
lugar_trabajoDomicilio_clientes	0.0276
	(0.0814)
lugar_trabajoCalle	0.0056
	(0.1117)
lugar_trabajoOtro_lugar	0.0020
	(0.2118)
lugar_trabajoOtros	-0.2066
	(0.6640)
tipoOtros	0.9582*
	(0.5510)
tipoActividad_familiar	1.1296**
	(0.5584)
tipoOtro_tipo	0.5348
	(0.5959)
finalizacionNo_tiene_finalizacion	-0.1297
	(0.0854)
finalizacionOtros	-0.0094
	(0.1021)
tiempo_busqueda1_3_meses	-0.2462

	log-odds
	(0.1889)
tiempo_busqueda3_6_meses	-0.1946 (0.2259)
tiempo_busqueda6_12_meses	0.0711 (0.2243)
tiempo_busquedaMas_1_anio	0.0999 (0.1852)
tiempo_busquedaOtros	0.5375 (1.1717)
tiempo_ocup_anterior1_3_meses	0.2602 (0.1712)
tiempo_ocup_anterior3_6_meses	0.4015* (0.2178)
tiempo_ocup_anterior6_12_meses	0.5055** (0.2374)
tiempo_ocup_anterior1_3_anios	0.8145*** (0.2247)
tiempo_ocup_anteriorMas_3_anios	-0.9665 (2.7442)
tiempo_ocup_anteriorOtros	-0.6769 (2.7401)
aglomeradoBahia_Blanca	-0.2832 (0.2062)
aglomeradoRosario	-0.1806 (0.1740)
aglomeradoSanta_Fe	-0.0385 (0.1834)
aglomeradoParana	-0.0452 (0.1967)
aglomeradoPosadas	0.4901*** (0.1794)
aglomeradoResistencia	1.5073***

	log-odds
	(0.1569)
aglomeradoCdro_Rivadavia	0.1944
	(0.1951)
aglomeradoMendoza	0.0710
	(0.1734)
aglomeradoCorrientes	0.4653***
	(0.1683)
aglomeradoCordoba	0.0033
	(0.1667)
aglomeradoConcordia	0.1296
	(0.1838)
aglomeradoFormosa	0.9171***
	(0.1736)
aglomeradoNeuquen	0.3865**
	(0.1906)
aglomeradoStgo_del_Estero	0.0425
	(0.1781)
aglomeradoJujuy	0.1789
	(0.1761)
aglomeradoRio_Gallegos	0.2996
	(0.1996)
aglomeradoCatamarca	0.3184*
	(0.1697)
aglomeradoSalta	-0.0289
	(0.1694)
aglomeradoLa_Rioja	-0.0267
	(0.1853)
aglomeradoSan_Luis	0.1092
	(0.1894)
aglomeradoSan_Juan	0.0529
	(0.1767)
aglomeradoTucuman	-0.2493

	log-odds
	(0.1699)
aglomeradoSanta_Rosa	-0.5273**
	(0.2507)
aglomeradoUshuaia	0.3822**
	(0.1941)
aglomeradoCABA	-0.2925
	(0.1927)
aglomeradoGBA	-0.2193
	(0.1484)
aglomeradoMar_del_Plata	-0.1332
	(0.1880)
aglomeradoRio_Cuarto	-0.3386
	(0.2115)
aglomeradoSan_Nicolas	0.1989
	(0.1890)
aglomeradoRawson	0.0377
	(0.2023)
aglomeradoViedma	0.2835
	(0.1986)
tasa_desoc	0.7481
	(0.9349)
tasa_empleo	1.6103
	(3.5201)
tasa_inactividad	1.4897
	(3.3989)
log_ingreso_real	-0.1350***
	(0.0362)
log_ipcf_real	0.1501
	(0.1958)
log_ingreso_otras_ocupaciones	-0.0760**
	(0.0305)
log_ingreso_nolaboral	0.3339***

	log-odds
	(0.0183)
log_ingreso_total_individual	-0.0229
	(0.0315)
lider	-0.0044
	(0.0031)
emae_orig	0.0079***
	(0.0029)
tcnm	0.0586
	(0.0683)
villa	-0.0224
	(0.1745)
calidad_viviendaParcialmente_insuficiente	-0.0875
	(0.1194)
calidad_viviendaInsuficiente	-0.1397
	(0.1256)
calidad_viviendaOtros	0.0788
	(0.1900)
fuentegto_trabajo	-0.1627
	(0.1137)
fuentegto_jub_pens	0.2675***
	(0.0507)
fuentegto_indemn_despido	-0.9263***
	(0.2771)
fuentegto_seguro_desemp	-1.1098***
	(0.2474)
fuentegto_subsidio	-0.2163***
	(0.0683)
fuentegto_alquiler	-0.4523***
	(0.1271)
fuentegto_otro_negocio	-0.7811
	(0.5183)
fuentegto_intereses_pf	-0.3308

	log-odds
	(0.4033)
fuentegto_ahorro	-0.0449
	(0.0561)
fuentegto_fliares	0.0051
	(0.0719)
fuentegto_banco	-0.0317
	(0.0722)
fuentegto_tarjcredito	0.0464
	(0.0448)
deccfr0	0.6591
	(0.6994)
deccfr2	0.2837***
	(0.0972)
deccfr3	0.1418
	(0.1218)
deccfr4	0.0993
	(0.1416)
deccfr5	0.1867
	(0.1583)
deccfr6	0.1965
	(0.1717)
deccfr7	0.2022
	(0.1884)
deccfr8	0.2051
	(0.2086)
deccfr9	0.2787
	(0.2315)
deccfr10	0.1067
	(0.2830)
mujer:sector_2S2	0.5686
	(0.4885)
mujer:sector_2S3	0.3803*

	log-odds
	(0.1980)
mujer:sector_2S4	-0.1573 (0.4530)
mujer:sector_2S5	0.1104 (0.4018)
mujer:sector_2S6	-0.1084 (0.2590)
mujer:sector_2S7	0.3473* (0.1917)
mujer:sector_2S8	0.2748 (0.2440)
mujer:sector_2S9	0.1048 (0.2216)
mujer:sector_2S10	0.2128 (0.2921)
mujer:sector_2S11	0.1408 (0.3138)
mujer:sector_2S12	0.0439 (0.3999)
mujer:sector_2S13	-0.2589 (0.2161)
mujer:sector_2S14	0.3684 (0.2697)
mujer:sector_2S15	-0.3012 (0.2108)
mujer:sector_2S16	-0.0391 (0.2307)
mujer:sector_2S17	0.3398 (0.2576)
mujer:sector_2S18	0.0473 (0.2420)
mujer:sector_2S19	-0.1230

	log-odds
	(0.2185)
mujer:sector_2S20	-0.0532 (0.3287)
mujer:sector_2S21	-0.6602 (0.5688)
mujer:sector_2S22	-0.3666 (0.7893)
mujer:miembros	-0.0038 (0.0176)
mujer:menor10_tot_hogar	0.1481*** (0.0531)
mujer:menor6_tot_hogar	0.0153 (0.0710)
mujer:menor1_tot_hogar	0.0740 (0.1230)
mujer:jefe_parejaOtros	-0.5821*** (0.0894)
mujer:jefe_parejaConyuge_pareja	-0.1448 (0.0931)
mujer:edad_2	-0.0006** (0.0003)
mujer:edad	0.0265 (0.0221)
mujer:casado	0.9073*** (0.0744)
mujer:educNivel_medio	-0.1275** (0.0575)
mujer:educNivel_alto	-0.1115 (0.1166)
mujer:cat_ocupOtros	10.7225 (324.7549)
mujer:cat_ocupPatron	0.2478

	log-odds
	(0.3405)
mujer:cat_ocupCuenta_propia_noprof	0.1501
	(0.1518)
mujer:cat_ocupFliar_s_remuneracion	-0.3274
	(0.3563)
mujer:cat_ocupCuenta_propia_prof	0.1236
	(0.2471)
mujer:formalOtros	-0.2215
	(0.2622)
mujer:formalInformal	0.3964***
	(0.0932)
mujer:ocup_desocupDesocup_ocup_anterior	0.0150
	(0.1385)
mujer:ocup_desocupOcupado	10.5293
	(324.7602)
mujer:extranjero	0.0389
	(0.1380)
mujer:tot_horas_os	-0.0058
	(0.0105)
mujer:tot_horas_op	-0.0026
	(0.0028)
mujer:intens_horariaOtros	-0.2065
	(1.4086)
mujer:intens_horariaSubocup_voluntaria	-0.2832***
	(0.0977)
mujer:intens_horariaSubocup_involuntaria	-0.1968*
	(0.1145)
mujer:intens_horariaSobre_ocup	0.0190
	(0.0890)
mujer:antigOtros	0.2498
	(0.4957)
mujer:antig3_6_meses	-0.3487**

	log-odds
	(0.1409)
mujer:antig6_12_mes	-0.1476 (0.1364)
mujer:antig1_5_anios	-0.1385 (0.1037)
mujer:antigmas_5_anios	-0.2640** (0.1083)
mujer:jerarqDireccion	-0.2111 (1.5282)
mujer:jerarqJefe	-0.5168 (1.5385)
mujer:jerarqTrabaj_asalariado	-0.3636 (0.2399)
mujer:jerarqOtros	1.3287 (2.6639)
mujer:tecno_ocupCon_maquina	0.1616 (0.1169)
mujer:tecno_ocupEquipos_informaticos	-0.0476 (0.1056)
mujer:tecno_ocupOtros	0.3305 (1.5044)
mujer:calificTecnico	0.1431 (0.1595)
mujer:calificOperativo	0.0815 (0.1632)
mujer:calificNo_calificado	0.1204 (0.1779)
mujer:calificOtros	-1.5594 (1.2772)
mujer:tam_establecimientoEntre_6_40	-0.1475 (0.0905)
mujer:tam_establecimientoMas_40	-0.1161

	log-odds
	(0.1091)
mujer:tam_establecimientoOtros	-0.0629 (0.1339)
mujer:tipo_viviendaDepartamento	0.0120 (0.0789)
mujer:tipo_viviendaPieza_inquilinato	-0.1360 (0.3797)
mujer:tipo_viviendaPieza_hotel_pension	0.2553 (0.9063)
mujer:tipo_viviendaLocal_noconstruido_p_habitacion	7.5095 (44.5501)
mujer:tipo_viviendaOtros	-0.2732 (0.4763)
mujer:lugar_trabajoPuesto_kiosco	-0.4196 (0.3932)
mujer:lugar_trabajoVehiculo	-0.5372 (0.4745)
mujer:lugar_trabajoVehiculo_transporte_personas	-0.0188 (0.2512)
mujer:lugar_trabajoObra_construccion	0.3690 (0.3854)
mujer:lugar_trabajoVivienda_propia	0.1221 (0.1323)
mujer:lugar_trabajoVivienda_socio_patron	0.1294 (0.4044)
mujer:lugar_trabajoDomicilio_clientes	0.2864*** (0.1029)
mujer:lugar_trabajoCalle	0.3216** (0.1353)
mujer:lugar_trabajoOtro_lugar	0.6291** (0.2609)
mujer:lugar_trabajoOtros	-0.6022

	log-odds
	(0.7246)
mujer:tipoOtros	-0.6902 (0.6239)
mujer:tipoActividad_familiar	-0.8271 (0.6316)
mujer:tipoOtro_tipo	-0.5922 (0.6829)
mujer:finalizacionNo_tiene_finalizacion	-0.1002 (0.1029)
mujer:finalizacionOtros	0.0863 (0.1220)
mujer:tiempo_búsqueda1_3_meses	0.3803* (0.2293)
mujer:tiempo_búsqueda3_6_meses	0.3160 (0.2700)
mujer:tiempo_búsqueda6_12_meses	0.0520 (0.2686)
mujer:tiempo_búsquedaMas_1_año	0.0387 (0.2255)
mujer:tiempo_búsquedaOtros	0.2858 (1.7070)
mujer:tiempo_ocup_anterior1_3_meses	-0.4017* (0.2119)
mujer:tiempo_ocup_anterior3_6_meses	-0.4045 (0.2648)
mujer:tiempo_ocup_anterior6_12_meses	-0.3226 (0.2843)
mujer:tiempo_ocup_anterior1_3_años	-0.5765** (0.2671)
mujer:tiempo_ocup_anteriorMas_3_años	-10.5655 (324.7557)
mujer:tiempo_ocup_anteriorOtros	-11.1625

	log-odds
	(324.7556)
mujer:aglomeradoBahia_Blanca	0.3689 (0.2369)
mujer:aglomeradoRosario	0.3780* (0.2012)
mujer:aglomeradoSanta_Fe	0.2865 (0.2136)
mujer:aglomeradoParana	0.2476 (0.2263)
mujer:aglomeradoPosadas	0.0357 (0.2107)
mujer:aglomeradoResistencia	-0.4980*** (0.1896)
mujer:aglomeradoCdro_Rivadavia	0.4354* (0.2275)
mujer:aglomeradoMendoza	0.3191 (0.2012)
mujer:aglomeradoCorrientes	-0.2705 (0.2000)
mujer:aglomeradoCordoba	0.2138 (0.1948)
mujer:aglomeradoConcordia	0.6150*** (0.2123)
mujer:aglomeradoFormosa	0.0572 (0.2080)
mujer:aglomeradoNeuquen	0.1749 (0.2223)
mujer:aglomeradoStgo_del_Estero	0.5807*** (0.2096)
mujer:aglomeradoJujuy	0.0296 (0.2064)
mujer:aglomeradoRio_Gallegos	-0.0641

	log-odds
	(0.2378)
mujer:aglomeradoCatamarca	-0.1199
	(0.2018)
mujer:aglomeradoSalta	0.0101
	(0.1983)
mujer:aglomeradoLa_Rioja	0.0367
	(0.2181)
mujer:aglomeradoSan_Luis	0.1773
	(0.2209)
mujer:aglomeradoSan_Juan	0.3042
	(0.2080)
mujer:aglomeradoTucuman	0.4783**
	(0.1969)
mujer:aglomeradoSanta_Rosa	0.1669
	(0.2879)
mujer:aglomeradoUshuaia	0.2011
	(0.2269)
mujer:aglomeradoCABA	0.4162*
	(0.2217)
mujer:aglomeradoGBA	0.4560***
	(0.1726)
mujer:aglomeradoMar_del_Plata	0.0275
	(0.2193)
mujer:aglomeradoRio_Cuarto	0.3315
	(0.2424)
mujer:aglomeradoSan_Nicolas	0.1680
	(0.2228)
mujer:aglomeradoRawson	0.1384
	(0.2362)
mujer:aglomeradoViedma	-0.1889
	(0.2333)
mujer:tasa_desoc	-1.8306*

	log-odds
	(1.0756)
mujer:tasa_empleo	-5.2944 (3.9912)
mujer:tasa_inactividad	-4.7848 (3.8482)
mujer:log_ingreso_real	-0.0322 (0.0418)
mujer:log_ipcf_real	-0.0080 (0.2262)
mujer:log_ingreso_otras_ocupaciones	-0.0410 (0.0348)
mujer:log_ingreso_nolaboral	-0.2702*** (0.0216)
mujer:log_ingreso_total_individual	-0.0283 (0.0366)
mujer:lider	0.0039 (0.0037)
mujer:emae_orig	-0.0023 (0.0034)
mujer:tcnm	-0.0869 (0.0808)
mujer:villa	0.1469 (0.1991)
mujer:calidad_viviendaParcialmente_insuficiente	0.1776 (0.1374)
mujer:calidad_viviendaInsuficiente	0.2433* (0.1446)
mujer:calidad_viviendaOtros	0.0443 (0.2214)
mujer:fuentegto_trabajo	0.0510 (0.1381)
mujer:fuentegto_jub_pens	-0.0731

	log-odds
	(0.0613)
mujer:fuentegto_indemn_despido	0.8288***
	(0.3165)
mujer:fuentegto_seguro_desemp	0.9982***
	(0.2841)
mujer:fuentegto_subsidio	0.1109
	(0.0806)
mujer:fuentegto_alquiler	0.5380***
	(0.1503)
mujer:fuentegto_otro_negocio	0.8967
	(0.5963)
mujer:fuentegto_intereses_pf	0.0209
	(0.5105)
mujer:fuentegto_ahorro	0.0165
	(0.0648)
mujer:fuentegto_fliares	-0.0196
	(0.0819)
mujer:fuentegto_banco	0.0207
	(0.0836)
mujer:fuentegto_tarjcredito	-0.0134
	(0.0522)
mujer:deccfr0	-0.2490
	(0.8169)
mujer:deccfr2	-0.2793**
	(0.1121)
mujer:deccfr3	-0.1072
	(0.1402)
mujer:deccfr4	-0.0062
	(0.1632)
mujer:deccfr5	-0.1695
	(0.1832)
mujer:deccfr6	-0.1432

	log-odds
	(0.1990)
mujer:deccfr7	-0.0844 (0.2186)
mujer:deccfr8	-0.1073 (0.2421)
mujer:deccfr9	-0.0480 (0.2690)
mujer:deccfr10	0.1323 (0.3274)
AIC	66803.0103
BIC	70150.0931
Log Likelihood	-33067.5051
Deviance	66135.0103
Num. obs.	166243

Fuente: Elaboración propia con base en la EPH 2004-2016. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

C.3 Regresión lasso

Tabla 6: Regresión logit. Selección de variables modelo con género sin interacciones

Predictores	Coefficientes
(Intercept)	-3.72
sector_2S3	0.03
sector_2S4	0.00
sector_2S6	-0.06
sector_2S7	0.10
sector_2S9	0.08
sector_2S12	0.10
sector_2S15	-0.31
sector_2S16	-0.32
sector_2S17	-0.17
sector_2S19	-0.06

miembros	0.02
menor10_tot_hogar	0.00
menor6_tot_hogar	0.03
mujer	1.04
jefe_parejaOtros	0.32
jefe_parejaConyuge_pareja	0.64
edad_2	0.00
casado	0.08
educNivel_medio	-0.03
educNivel_alto	-0.55
cat_ocupOtros	0.63
cat_ocupCuenta_propia_noprof	0.29
formalOtros	0.29
formalInformal	0.78
tot_horas_os	-0.02
tot_horas_op	-0.02
intens_horariaSubocup_voluntaria	0.13
intens_horariaSobre_ocup	0.04
antigOtros	-0.08
antig3_6_meses	0.23
antig6_12_meses	0.20
antigmas_5_anios	-0.30
jerarqTrabaj_asalariado	-0.10
tecno_ocupEquipos_informaticos	-0.12
calificTecnico	-0.02
calificOperativo	0.00
tam_establecimientoMas_40	-0.11
tipo_viviendaDepartamento	-0.02
tipo_viviendaLocal_noconstruido_p_habitacion	-0.16
lugar_trabajoPuesto_kiosco	0.09
lugar_trabajoVivienda_propia	0.49
lugar_trabajoVivienda_socio_patron	0.17
lugar_trabajoDomicilio_clientes	0.07

lugar_trabajoCalle	0.10
lugar_trabajoOtro_lugar	0.14
lugar_trabajoOtros	-0.05
tipoActividad_familiar	0.06
tipoOtro_tipo	-0.12
finalizacionNo_tiene_finalizacion	-0.20
tiempo_búsqueda6_12_meses	0.03
tiempo_búsquedaMas_1_año	0.03
tiempo_ocup_anterior3_6_meses	0.01
tiempo_ocup_anterior6_12_meses	0.19
tiempo_ocup_anterior1_3_años	0.39
tiempo_ocup_anteriorMas_3_años	0.03
aglomeradoBahia_Blanca	-0.05
aglomeradoPosadas	0.19
aglomeradoResistencia	0.86
aglomeradoCdro_Rivadavia	0.25
aglomeradoMendoza	0.03
aglomeradoConcordia	0.30
aglomeradoFormosa	0.64
aglomeradoNeuquen	0.19
aglomeradoStgo_del_Estero	0.16
aglomeradoSalta	-0.08
aglomeradoLa_Rioja	-0.04
aglomeradoSanta_Rosa	-0.37
aglomeradoUshuaia	0.21
aglomeradoCABA	-0.09
aglomeradoMar_del_Plata	-0.14
aglomeradoRio_Cuarto	-0.12
aglomeradoSan_Nicolas	0.05
tasa_inactividad	0.02
log_ingreso_real	-0.18
log_ipcf_real	0.13
log_ingreso_otras_ocupaciones	-0.12

log_ingreso_nolaboral	0.12
log_ingreso_total_individual	-0.05
emae_orig	0.00
calidad_viviendaInsuficiente	0.00
fuentegto_jub_pens	0.24
fuentegto_indemn_despido	-0.04
fuentegto_seguro_desemp	-0.20
fuentegto_subsidio	-0.15
deccfr0	0.11
deccfr9	0.02

Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004 - 2017

Tabla 7: Regresión logit. Selección de variables modelo con interacciones

Predictores	Coefficientes
(Intercept)	-3.24
sector_2S4	0.00
sector_2S9	0.03
sector_2S12	0.07
sector_2S15	-0.04
sector_2S16	-0.10
sector_2S17	-0.07
menor10_tot_hogar	-0.03
menor1_tot_hogar	-0.00
jefe_parejaOtros	0.32
jefe_parejaConyuge_pareja	0.17
edad_2	0.00
casado	-0.25
educNivel_medio	-0.03
educNivel_alto	-0.51
cat_ocupOtros	0.67
cat_ocupCuenta_propia_noprof	0.25
formalInformal	0.43
tot_horas_os	-0.01

tot_horas_op	-0.02
intens_horariaSubocup_voluntaria	0.12
antig3_6_meses	0.23
antig6_12_meses	0.12
antigmas_5_anios	-0.22
tecno_ocupEquipos_informaticos	-0.04
tam_establecimientoMas_40	-0.13
tipo_viviendaDepartamento	-0.01
lugar_trabajoPuesto_kiosco	0.09
lugar_trabajoVivienda_propia	0.31
lugar_trabajoVivienda_socio_patron	0.05
tipoActividad_familiar	0.01
tipoOtro_tipo	-0.07
finalizacionNo_tiene_finalizacion	-0.23
tiempo_búsqueda6_12_meses	0.03
tiempo_búsquedaMas_1_año	0.05
tiempo_ocup_anterior1_3_meses	-0.02
tiempo_ocup_anterior6_12_meses	0.12
tiempo_ocup_anterior1_3_años	0.30
aglomeradoBahia_Blanca	-0.04
aglomeradoPosadas	0.19
aglomeradoResistencia	0.99
aglomeradoFormosa	0.66
aglomeradoNeuquen	0.08
aglomeradoCatamarca	0.01
aglomeradoSanta_Rosa	-0.34
aglomeradoCABA	-0.08
aglomeradoMar_del_Plata	-0.04
aglomeradoRio_Cuarto	-0.10
tasa_inactividad	0.01
log_ingreso_real	-0.16
log_ipcf_real	0.09
log_ingreso_otras_ocupaciones	-0.12

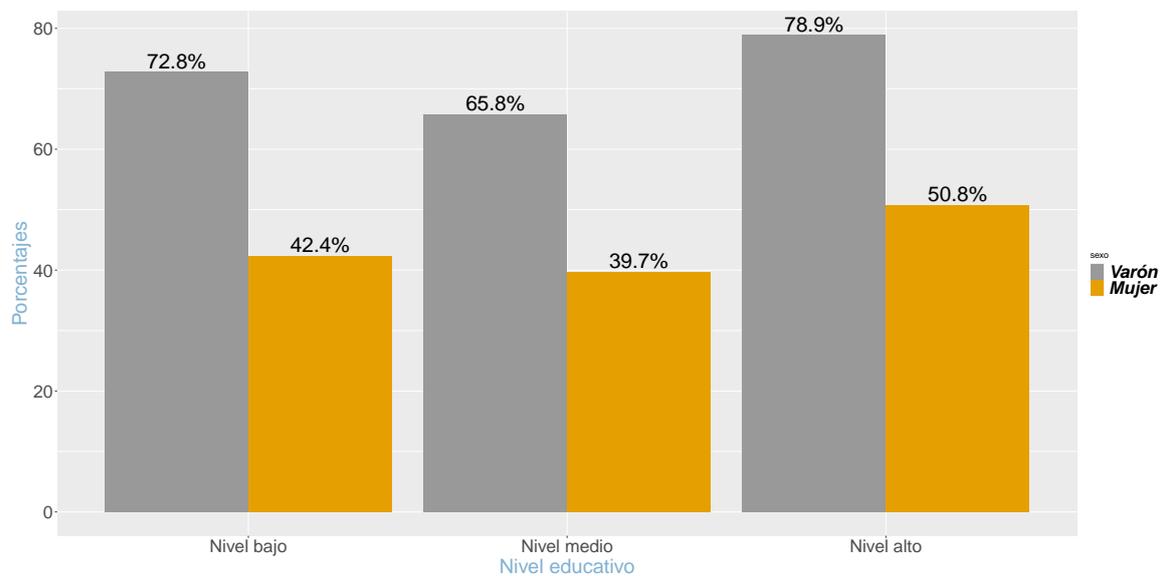
log_ingreso_nolaboral	0.23
log_ingreso_total_individual	-0.05
emae_orig	0.00
fuentegto_jub_pens	0.31
fuentegto_indemn_despido	-0.09
fuentegto_seguro_desemp	-0.26
fuentegto_subsidio	-0.09
mujer:sector_2S3	0.31
mujer:sector_2S7	0.30
mujer:sector_2S8	0.08
mujer:sector_2S9	0.19
mujer:sector_2S10	0.02
mujer:sector_2S12	0.17
mujer:sector_2S14	0.24
mujer:sector_2S15	-0.32
mujer:sector_2S16	-0.06
mujer:sector_2S18	0.06
mujer:miembros	0.01
mujer:menor10_tot_hogar	0.13
mujer:menor6_tot_hogar	0.05
mujer:jefe_parejaConyuge_pareja	0.13
mujer:casado	0.84
mujer:cat_ocupCuenta_propia_noprof	0.09
mujer:cat_ocupFluar_s_remuneracion	-0.07
mujer:formalOtros	0.44
mujer:formalInformal	0.50
mujer:tot_horas_os	-0.01
mujer:intens_horariaOtros	0.05
mujer:intens_horariaSobre_ocup	0.15
mujer:antig6_12_meses	0.10
mujer:antigmas_5_anios	-0.09
mujer:tecno_ocupEquipos_informaticos	-0.00
mujer:tecno_ocupOtros	0.11

mujer:calificOperativo	0.01
mujer:calificNo_calificado	0.01
mujer:tipo_viviendaLocal_noconstruido_p_habitacion	-0.25
mujer:lugar_trabajoVehiculo_transporte_personas	0.20
mujer:lugar_trabajoVivienda_propia	0.18
mujer:lugar_trabajoVivienda_socio_patron	0.26
mujer:lugar_trabajoDomicilio_clientes	0.30
mujer:lugar_trabajoCalle	0.29
mujer:lugar_trabajoOtro_lugar	0.45
mujer:tiempo_búsqueda1_3_meses	0.04
mujer:tiempo_ocup_anterior3_6_meses	0.04
mujer:tiempo_ocup_anterior6_12_meses	0.05
mujer:aglomeradoResistencia	-0.14
mujer:aglomeradoCdro_Rivadavia	0.41
mujer:aglomeradoMendoza	0.08
mujer:aglomeradoConcordia	0.46
mujer:aglomeradoNeuquen	0.18
mujer:aglomeradoStgo_del_Estero	0.31
mujer:aglomeradoSalta	-0.17
mujer:aglomeradoLa_Rioja	-0.07
mujer:aglomeradoSan_Juan	0.04
mujer:aglomeradoUshuaia	0.30
mujer:aglomeradoGBA	0.01
mujer:aglomeradoMar_del_Plata	-0.14
mujer:aglomeradoSan_Nicolas	0.08
mujer:log_ingreso_nolaboral	-0.14
mujer:villa	0.09
mujer:calidad_viviendaInsuficiente	0.03
mujer:fuentegto_jub_pens	-0.10
mujer:fuentegto_tarjcredito	0.01
mujer:deccfr0	0.15
mujer:deccfr9	0.04
mujer:deccfr10	0.00

Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004 - 2017

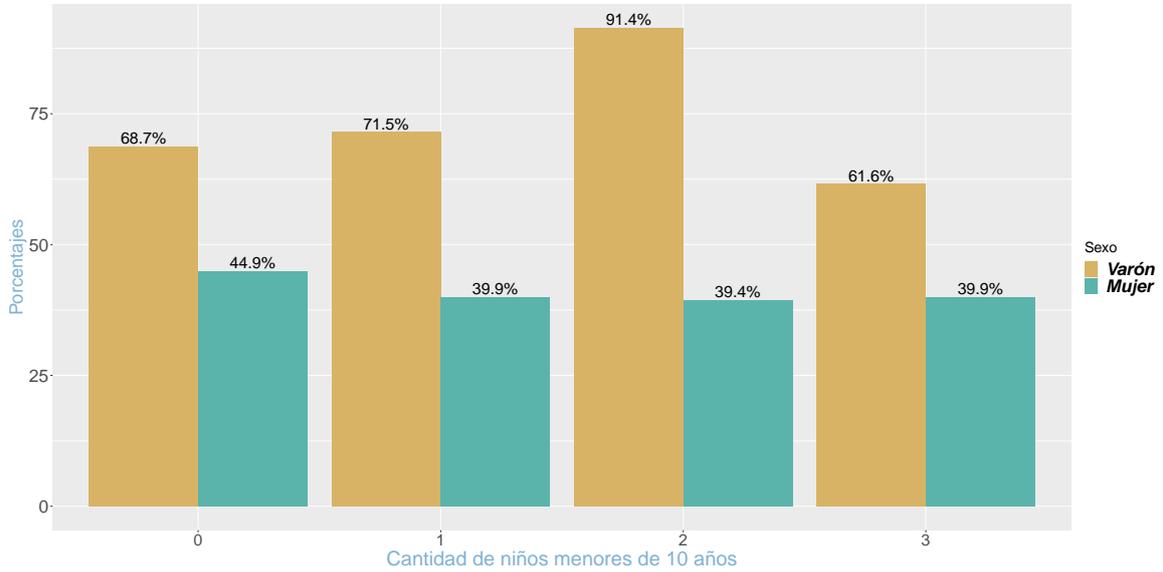
C.4 Post-estimación: transición a la actividad laboral

Figura 32: Proporción de inactivos que pasan a la actividad según género y nivel educativo



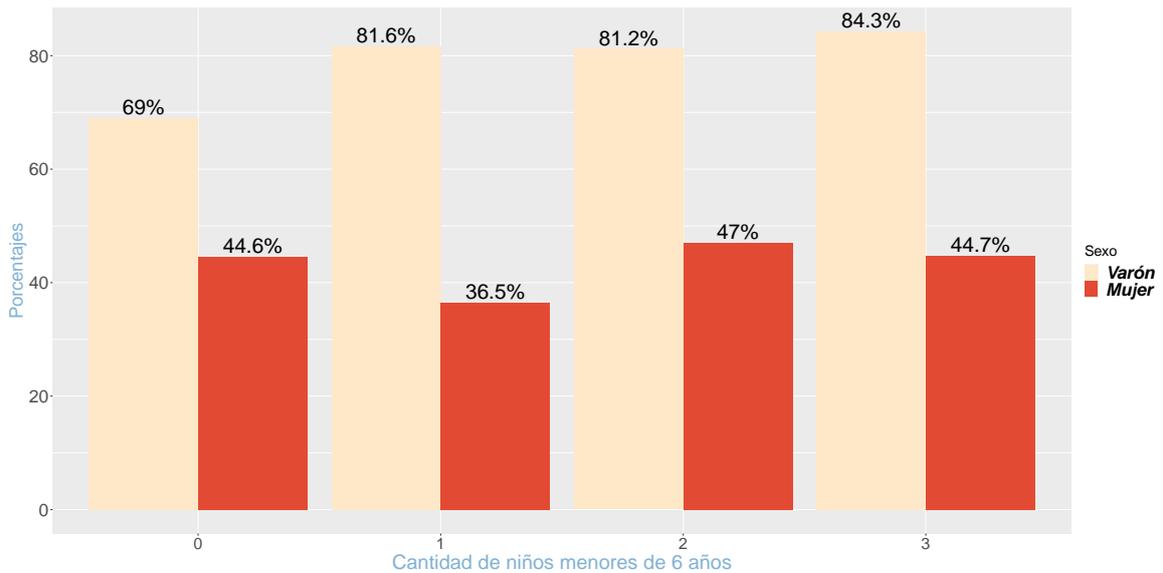
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 33: Proporción de inactivos que pasan a la actividad según género y presencia de menores de 10 años



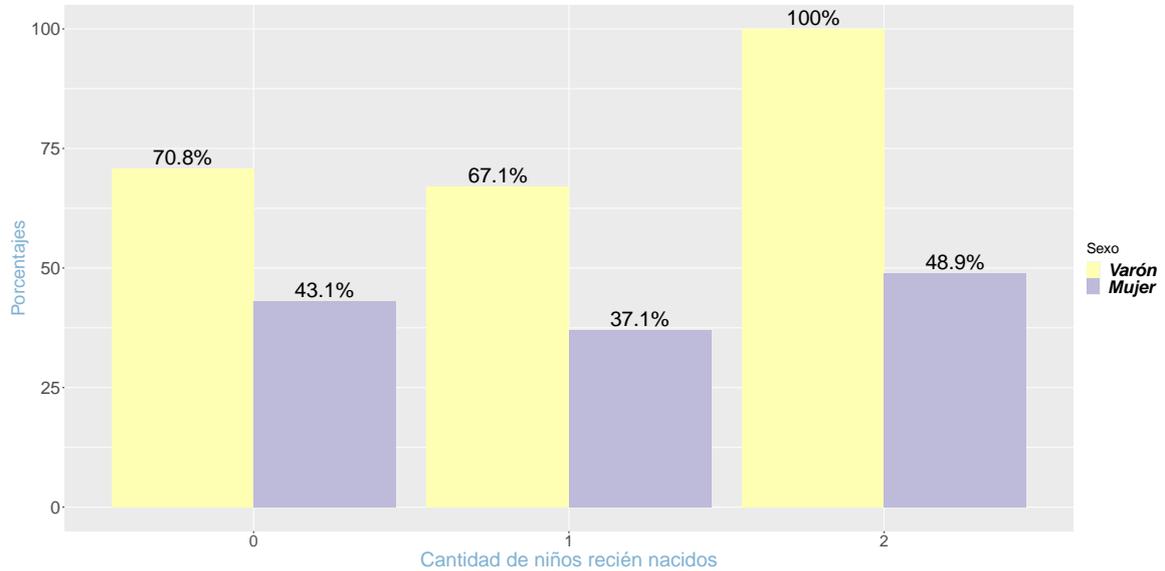
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 34: Proporción de inactivos que pasan a la actividad según género y presencia de menores de 6 años



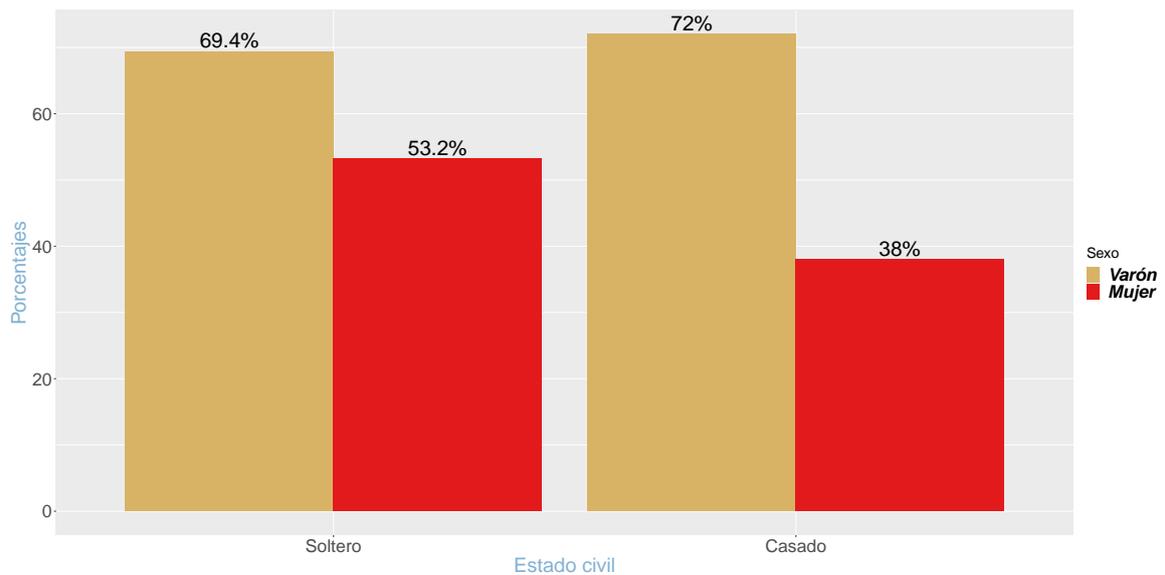
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 35: Proporción de inactivos que pasan a la actividad según género y presencia de recién nacidos



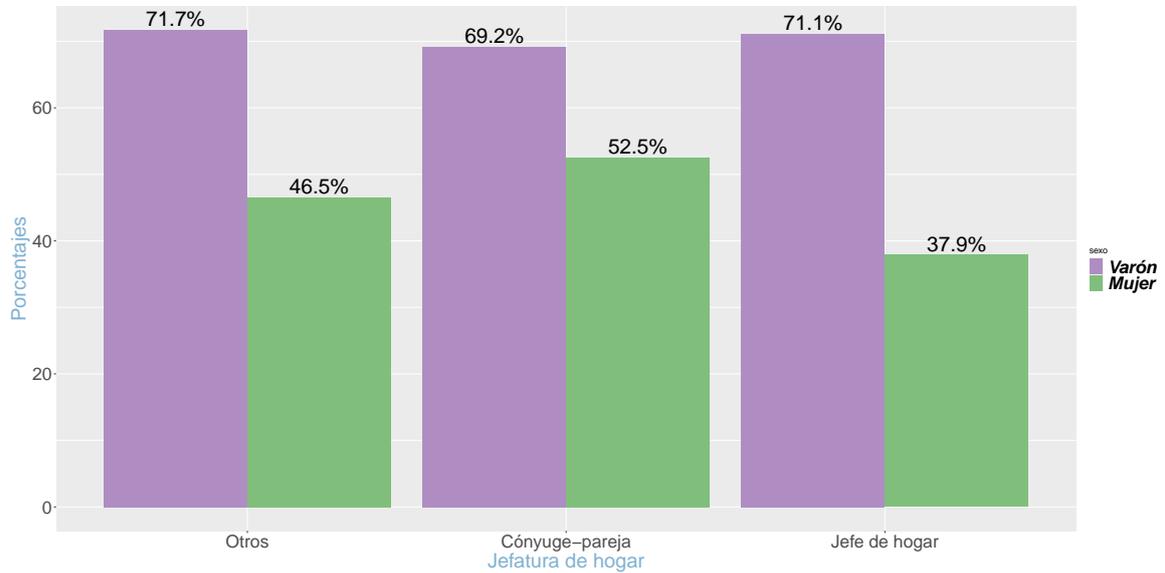
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 36: Proporción de inactivos que pasan a la actividad según género y estado civil



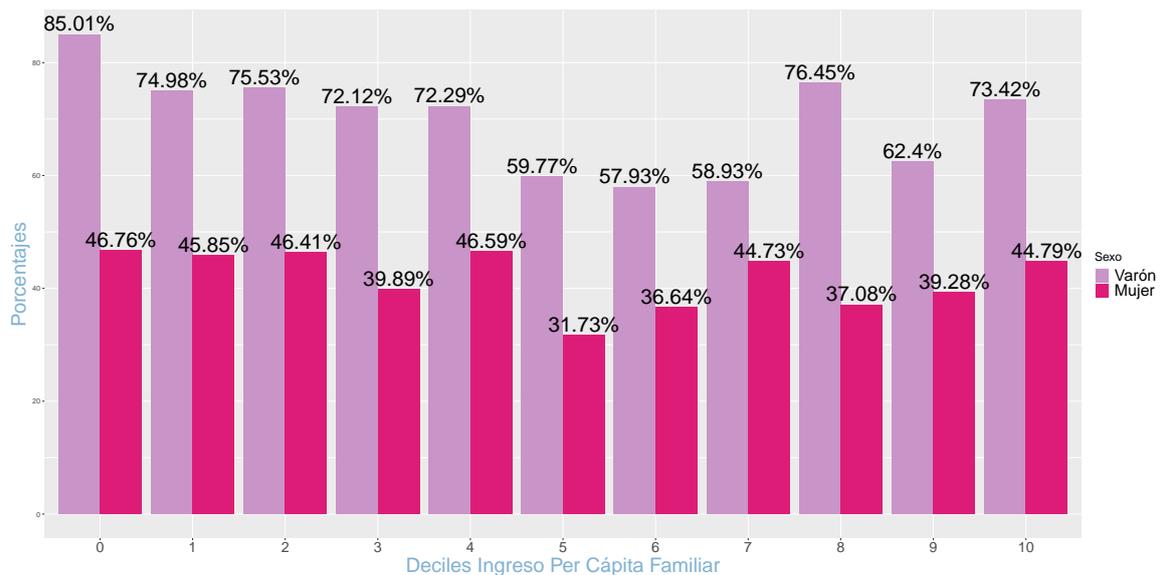
Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 37: Proporción de inactivos que pasan a la actividad según género y jefatura del hogar



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016

Figura 38: Proporción de inactivos que pasan a la actividad según género y decil del ingreso per cápita familiar



Fuente: Elaboración propia con base en la Encuesta Permanente de Hogares 2004-2016