2 Fraude en subvenciones públicas: pilotar un modelo de riesgo basado en datos en España

Este capítulo presenta una prueba de concepto para un modelo de riesgo que la Intervención General de la Administración del Estado (IGAE) de España puede emplear para evaluar los riesgos de fraude y detectar posibles casos de fraude. El capítulo presenta una descripción general de la metodología de aprendizaje automático que subyace en el modelo de riesgo, así como un relato detallado de cómo se construyó el modelo, basado en datos que están fácilmente disponibles para la IGAE. El capítulo concluye con una exposición de los resultados del modelo y recomendaciones para que la IGAE se base en la prueba de concepto.

Introducción

Los marcos de evaluación de riesgos de fraude basados en datos pueden tener múltiples usos, entre los cuales es fundamental identificar las prioridades de investigación. Cuando los recursos de investigación son escasos, y es probable que una selección aleatoria de casos para la investigación arroje una tasa de éxito baja (por ejemplo, porque el fraude es poco frecuente en la población objetivo), una selección de casos basada en riesgos puede reportar beneficios significativos. Con este fin, una calificación de riesgo asignada a todos los casos potencialmente investigados puede contribuir a priorizar los casos que deben investigarse. Por lo general, esto no implica una automatización completa de la selección de casos, pero sí ofrece una aportación crucial en el proceso de toma de decisiones de la organización.

Para que una evaluación de riesgos a gran escala reporte beneficios, debe ser lo suficientemente precisa como para ser utilizada para operaciones u organizaciones de calificación de riesgos de manera continua, incluidos los casos nuevos. En general, las calificaciones de riesgo pueden definirse como válidas para dichos propósitos, ya sea definiendo explícitamente los factores de riesgo a partir de relaciones conocidas y descripciones de riesgo (por ejemplo, el propietario último de la organización receptora de la subvención reside en un paraíso fiscal), o definiendo la combinación de características de riesgo a través de medios estadísticos, incluido el aprendizaje automático de investigaciones anteriores. De cualquier manera, lo que es crucial es que el modelo de riesgo no solo tenga en cuenta los casos fraudulentos conocidos y sus características, sino que también tenga en cuenta las características de un grupo mucho mayor de casos que no hayan sido investigados, por lo que se desconoce su situación respecto al fraude. En resumen, la validación de indicadores y la mejora continua son de crucial importancia, como se muestra en este capítulo.

Cuando se desarrollan nuevos enfoques analíticos, la comprensión y el aprendizaje suelen derivarse de la práctica. Este es el motivo por el que muchas instituciones de auditoría, por ejemplo, han creado «Laboratorios de innovación» y comunidades internas de práctica para probar y experimentar con nuevas técnicas de auditoría, análisis y tecnologías y uso de los datos. Este enfoque incremental permite a los organismos de auditoría y control asumir riesgos medidos y contener los costes, antes de ampliar o reducir las iniciativas piloto. Con este espíritu y en respuesta al interés de la Intervención General de la Administración del Estado (IGAE) en fortalecer su uso de datos para detectar riesgos de fraude en subvenciones, este capítulo presenta una prueba de concepto para un modelo de riesgo basado en datos para que la IGAE lo adopte en parte o en su totalidad.

La metodología de este capítulo tiene como objetivo hacer uso de los datos que ya estaban a disposición de la IGAE, incluida la Base de Datos Nacional de Subvenciones (BDNS) y, al hacerlo, implícitamente tiene en cuenta el contexto de la IGAE. Como se indica en el Capítulo 1, como cualquier inversión para mejorar la gobernanza de datos, la gestión de datos o el análisis, este enfoque puede requerir inversiones en aptitudes y habilidades digitales. Por este motivo, el capítulo proporciona una descripción detallada de todas las etapas de la metodología y su desarrollo para respaldar la propia evaluación de la IGAE de lo que es capaz de hacer con sus recursos y habilidades internas existentes. Además, el proceso de desarrollo de la prueba de concepto para el modelo de riesgo ha derivado en diversos descubrimientos y en la identificación de áreas de mejora, que se abordan en la sección de resultados.

Aspectos generales del modelo de aprendizaje automático

Una breve introducción al aprendizaje automático para las evaluaciones de riesgos

El enfoque actual de la IGAE para evaluar riesgos de fraude, detallado en el Capítulo 1, se describe en su Plan de Auditorías y Control Financiero de Subvenciones y Ayudas Públicas para 2021. La IGAE tiene en cuenta el importe de la subvención, los niveles previos de fraude y otros factores cualitativos, como los

procedimientos de justificación y verificación. El modelo de aprendizaje automático descrito en este capítulo avanza en un enfoque más basado en datos, que puede complementar los procesos existentes de la IGAE. En realidad, dadas las limitaciones de recursos, la IGAE solo puede realizar un número finito de actividades de control en un año. La metodología de aprendizaje automático descrita en este capítulo no debe sustituir el criterio del auditor. Por ejemplo, el modelo puede detectar casos de posible fraude, pero el auditor también necesitará evaluar cuál de estos casos es el más rentable en términos de actividades de investigación o control adicionales. Teniendo en cuenta este matiz, el modelo puede ser una aportación útil para las decisiones de los auditores y ayudar a la IGAE a dirigir sus recursos de manera más eficaz.

El modelo de riesgo desarrollado para apoyar la IGAE se basa en una metodología de *random forest*. Los *random forests* son un método de aprendizaje automático supervisado que predice el resultado mediante la formación de varios árboles de decisión con características determinadas (Breiman, 2001[1]). Es especialmente adecuado para conjuntos de datos con una gran cantidad de variables explicativas o indicadores de riesgo potencial. Al utilizar *random forests*, es posible incluir una amplia lista de factores explicativos de diferentes tipos (numéricos y categóricos).

Selección de la metodología

Para analizar los datos utilizando métodos de aprendizaje automático como *random forests*, el conjunto de datos se limpia previamente, eliminando los valores omitidos y las variables que carecen de varianza (es decir, donde las variables toman casi siempre el mismo valor en todo el conjunto de datos). Los *random forests* permiten trabajar con una gran cantidad de observaciones y variables, realizar entrenamiento de algoritmos en una muestra reducida y equilibrada, y probar modelos en una muestra reservada. Los algoritmos de *Random Forest* son sensibles a los valores omitidos. Por esta razón, se descartaron las variables con altos índices de omisión. El método también es sensible al desequilibrio en la variable dependiente (es decir, sancionado frente a no sancionado), como se describe a continuación. En general, el enfoque se puede dividir en los siguientes pasos:

1. Identificar qué beneficiarios fueron sancionados y luego marcar todas las concesiones a las organizaciones sancionadas en los últimos 2 ó 3 años antes de las sanciones. En este período, es muy probable que se haya producido una actividad fraudulenta probada. Esto da un conjunto completo de casos positivos probados (concesiones sancionadas); sin embargo, deja una muestra muy grande de casos sin etiquetar (no sancionados). Algunos de estos casos probablemente deberían haber sido sancionados, pero no fueron investigados, y otros son casos negativos reales en los que no habría habido sanción incluso si se hubieran investigado. En otras palabras, el conjunto de datos está muy desequilibrado. En la mayoría de los casos, se desconoce si la concesión no fue sancionada porque no fue investigada o porque no se descubrieron infracciones. Por tanto, la mayoría de las observaciones no son positivas ni negativas, sino que no están etiquetadas.
2. Elegir el método que se adapte al problema particular de los datos, es decir, una muestra desequilibrada y la presencia de una submuestra grande sin etiquetar. Para estos fines, se aplica un modelo de insaculación (*bagging)* positivo sin etiquetar (PU). Este método de aprendizaje automático permite entrenar un modelo en muestras aleatorias de observaciones, tanto positivas como sin etiquetar, para asignar un estado probablemente negativo (no sancionado) y un estado probablemente positivo (sancionado) a los casos sin etiquetar. El Recuadro 2.1 proporciona información adicional sobre la insaculación (*bagging)* de PU y los modelos de *random forest.*
3. Una vez asignadas las etiquetas, utilizar el conjunto de datos reetiquetado para entrenar al modelo, e identificar los factores que influyen en la probabilidad de ser sancionado. La influencia puede ser tanto positiva como negativa. Luego, el modelo calcula la probabilidad de que cada concesión sea sancionada por cualquier número de observaciones.

4. Una vez que el modelo esté entrenado y logre una precisión suficiente, aplicarlo al conjunto de datos completo de concesiones, para predecir una calificación de riesgo de fraude para todas las observaciones.[1]

Recuadro 2.1. Descripción general del aprendizaje y la insaculación de positivos sin etiquetar

El aprendizaje positivo sin etiquetar (PU) es una técnica de aprendizaje automático semisupervisada que permite trabajar con datos muy desequilibrados (Elkan and Noto, 2008[2]). El aprendizaje PU puede utilizarse en casos en los que la mayoría de las observaciones disponibles pertenecen a casos sin etiquetar. Esto incluye situaciones en las que una variable binaria (es decir, valores de 1 y 0) tiene observaciones positivas (1) que aparecen solo en caso de tratamiento, y cuando se desconoce si los casos negativos restantes (0) fueron tratados, pero siguieron siendo negativos, o no fueron tratados de ninguna manera. El aprendizaje PU observa todos los casos positivos y negativos, identifica las características más típicas de cada uno, y vuelve a etiquetar las observaciones como corresponde.

Un enfoque de *insaculación* PU consta de varios pasos (Li and Hua, 2014[3]). Primero, implica construir un clasificador analizando la variedad y combinación de factores que influyen en los resultados positivos y negativos. Para construir un clasificador, se crea un subconjunto de datos, que consta de todos los casos positivos y una muestra aleatoria de los no etiquetados. Este clasificador se aplica además al resto de casos sin etiquetar, para asignar las puntuaciones de probabilidad para el resto de las observaciones. Cada paso se repite varias veces, y luego se calcula la puntuación media recibida por cada observación.

Después de volver a etiquetar todas las observaciones, se divide en muestras de entrenamiento y prueba. La proporción de la división es flexible, pero suele estar entre el 60 % y el 70 % para la muestra de entrenamiento, y entre el 30 % y el 40 % para la muestra de prueba. Después, se aplica el método de *random forests* al conjunto de datos de entrenamiento. Los parámetros del modelo se pueden especificar manualmente, incluido el número de árboles, el número máximo de características en cada árbol individual y el tamaño de los nodos terminales. La elección de los parámetros depende del tamaño total del conjunto de datos, es decir, el número de observaciones e indicadores incluidos en el modelo. Después de aplicar el método de *random forests* a la muestra de entrenamiento, se pueden predecir las probabilidades de salida para el resto de los datos.

Además, para identificar el impacto de cada indicador, los valores SHAP (explicaciones aditivas de Shapley) se pueden calcular una vez que se construye el modelo. Los valores SHAP muestran cuánto y en qué dirección (positiva o negativa) se ha modificado la salida prevista para un indicador determinado. Para estimar el ajuste del modelo, deben calcularse parámetros tales como exactitud, repetición y precisión. Mediante ellos se calcula el número de predicciones de calificación correctas en términos absolutos o relativos.

Fuente: (Mordelet and Vert, 2014[4])

Consideración de puntos fuertes, puntos débiles y supuestos

La validez del análisis depende de dos factores: la calidad del conjunto de datos de entrenamiento y la disponibilidad de las características relevantes de la concesión, la ayuda y el beneficiario. En primer lugar, el principal indicador que diferencia los casos fraudulentos de los no fraudulentos es la presencia de sanciones. Para que el aprendizaje positivo sin etiquetar genere resultados válidos, se ha supuesto que los casos positivos se seleccionaron al azar, por lo que son una muestra representativa de todos los casos

positivos. Esto también implica que si la muestra de sanciones observadas no recoge algunos esquemas típicos de fraude (es decir, ni siquiera se encuentra un ejemplo entre los casos de sanciones observados), el modelo de aprendizaje automático no capturará dichos tipos de fraude, y por tanto, estará sesgado. De manera similar, si los casos que se seleccionaron siguiendo una variable en particular – por ejemplo, el tamaño del beneficiario -, el modelo sobreestimará la importancia de dicha variable en la predicción del riesgo. En otras palabras, el aprendizaje automático supervisado utiliza la información proporcionada por casos probados. Por tanto, si hay un sesgo en la muestra de concesiones sancionadas, se replicará en el proceso de predicción.

En segundo lugar, el modelo de aprendizaje automático solo puede conocer las características del fraude que capturan los datos. La presencia de ciertos indicadores en el conjunto de datos influye en el poder predictivo del modelo: si faltan algunos indicadores cruciales en los datos, el modelo no los tendrá en cuenta. Las características o indicadores que faltan también implican que la lista final de indicadores influyentes puede estar sesgada, exagerando la importancia de aquellas características que están correlacionadas con características influyentes pero no observadas (por ejemplo, si se encuentra que una región en particular tiene un mayor riesgo, en realidad, puede significar que algunas entidades en esa región tienen características de riesgo, como vínculos con políticos corruptos, y no que la región misma, su cultura, estructuras administrativas, etc., sean más propensas al fraude). Sin embargo, el método de aprendizaje automático elegido basado en *random forests* es especialmente adecuado para grandes conjuntos de datos con una gran cantidad de variables explicativas o indicadores de riesgo potencial (James et al., 2015[5]) Es posible incluir una amplia lista de factores explicativos de diferentes tipos (numéricos y categóricos).

Desarrollo de una prueba de concepto para un modelo de riesgo basado en datos

Identificar fuentes de datos y variables relevantes para evaluar riesgos de fraude de subvenciones

Los datos proporcionados por la IGAE constan de 17 conjuntos de datos que abarcan diferentes bloques de información sobre concesiones, terceros, proyectos, subvenciones y beneficiarios. Podrían agruparse en tres categorías principales.

- La primera categoría consta de siete conjuntos de datos que abarcan información sobre la convocatoria, como ubicación, tipo de actividad económica, objetivos e instrumentos.[2]
- La segunda categoría abarca información sobre concesiones, incluida información sobre reintegros, proyectos, devoluciones y detalles de las concesiones a beneficiarios.[3]
- La tercera categoría incluye conjuntos de datos que abarcan información sobre los propios beneficiarios, que pueden incluir una variedad de actores responsables de implantar un proyecto (por ejemplo, una entidad pública, contratista o subcontratista), si un beneficiario fue sancionado o inhabilitado, así como el tipo de actividad económica, ubicación, etc.[4]

En total, estos conjuntos de datos constan de alrededor de 100 variables que cubren detalles de las concesiones (importe, fecha de resolución, tipo de actividad económica, etc.), convocatorias de ayudas (publicidad, tipo de apoyo económico, base normativa, etc.) y datos de terceros (ubicación, naturaleza jurídica, actividades económicas, etc.). El período cubierto es 2018-2020.

Los tres grupos de conjuntos de datos presentan diferentes niveles de datos: la primera categoría abarca información de convocatorias y cada convocatoria puede abarcar varias concesiones. La segunda categoría incluye el nivel de concesión y puede vincularse al conjunto de datos principal BDNS_CONCESIONES mediante ID de concesiones únicas. Por último, la última categoría es sobre

beneficiarios y el mismo beneficiario puede recibir varias concesiones. Por tanto, con el fin de fusionar todos los conjuntos de datos entre sí, el nivel de concesión se utilizó como unidad principal de análisis, proporcionando ID únicas.

La lista de variables relevantes para la evaluación de riesgos de fraude podría dividirse en indicadores de antecedentes e indicadores de riesgo. Se necesitan indicadores de antecedentes para describir las características específicas de las convocatorias, los concedentes, los beneficiarios y terceros que están potencialmente asociados a las sanciones. Los indicadores de riesgo se refieren a determinadas fases de publicidad, selección, ejecución y seguimiento de subvenciones. La Tabla 2.1 muestra la lista completa de indicadores de antecedentes, la Tabla 2.2 muestra los indicadores de riesgo que podrían extraerse de los conjuntos de datos de la IGAE.

Tabla 2.1. Indicadores de antecedentes

Grupo de indicadores	Nombre del indicador	Código de variables	Encabezado de variables
Concedente	Organizador de convocatorias	CON 705; CON 710; CSU 100	Entidad organizadora; Entidad concedente; Entidad concedente
Concesionario	Beneficiario	CON 580; CSU 120	Tipos de beneficiario; ID del beneficiario
	ID de la concesión	PRO 110; PAG 100; DEV 100; REI 100	Identificación de la concesión
	ID del proyecto	PRO 130; EJE 110	Identificación del proyecto
	Descripción del proyecto	PRO 210	Descripción del proyecto
	localización del proyecto	PRO 260	Región geográfica (proyecto)
	Id de la ejecución	EJE 120	Identificación del ejecutor
	año	EJE 130	Año de ejecución
	ID descalificada	INH 100	ID descalificada
	Fecha de inhabilitación	INH 210	Fecha de descalificación
	Entidad inhabilitadora	INH 220	Identificar el origen administrativo o judicial de la entidad incapacitante
	Período de inhabilitación	INH 230; INH 240	Fecha de inicio de la inhabilitación; Fecha de finalización de la inhabilitación
Subvención	valor de la subvención	CSU 220; EJE 210	Importe de la subvención; Importe de la subvención a la entidad ejecutora por año
	base reguladora de la subvención	CON 250; CON 260	Descripción BBBR; URL BBBR
	Identificación de la convocatoria	CON 290	Título de la convocatoria
	publicación de la convocatoria	CON 300; CON 310	Enviado para publicación; Fuente oficial
		CON 335	Título en español, Texto en español
		CON 335; CON 340	Fecha de publicación; Enlace a la publicación
	fecha de la firma y lugar	CON 351; CON 352	Fecha de la firma; Ubicación de la firma
	solicitud	CON 440; CON 460	Fecha de inicio; Fecha de finalización
	ayuda estatal	CON 490; CON 495; CON 515	Condición de la ayuda estatal; autorización de la ayuda; Identificador de la ayuda de la UE
	sectores de convocatorias	CON 550	Sectores de Economía
	ubicación de la convocatoria	CON 570	Regiones geográficas
	plazos	CON 600	Plazo para justificar la concesión
	subvención nominativa	CON 610	Concesión de carácter nominativo
	Fondos de la UE	CON 690	Importe de financiación del fondo de la UE
	normativa	CSU 110	ID de la normativa
	fecha de pago	PÁG 210	Fecha de pago
	importe pagado	PÁG 220	Importe de pago
	retención	PÁG 230	Retención de impuestos
	fecha de devolución	DEV 210	Fechad e devolución

Grupo de indicadores	Nombre del indicador	Código de variables	Encabezado de variables
	importe de devolución	DEV 220	Importe de devolución
	tipo de interés	DEV 230	Importe del tipo de interés
	fecha de reembolso	REI 210	Fecha de reembolso
	motivo de reembolso	REI 220	El motivo del reembolso
	importe reembolsado	REI 230	El importe del reembolso
Tercero	país	TER 100; TER 250	País del tercero; País del domicilio
	id	TER 110	ID del tercero
	nombre	TER 240	Nombre del tercero; Nombre comercial del tercero
	apellido	TER 210	Primer apellido del tercero; Segundo apellido del tercero
	dirección	TER 252; TER 254; TER 256; TER 258; TER 310	Dirección del domicilio; Código Postal; Municipio; Provincia; Región
	tipo	TER 280; TER 290	Condición jurídica; Tipo de tercero
	actividad	TER 320	Sector de Economía

Fuente: Autor

Tabla 2.2. Indicadores de riesgo

Fase 2	Nombre del indicador	Definición del indicador	Variables (código)	Variables (encabezado)
Competencia	Falta de publicidad	No hay publicidad electrónica adecuada del programa de subvenciones	CON 310; CON 420; CON 620	Fuente oficial; Solicitud abierta; Condiciones de publicidad de la concesión
Selección	Proceso de selección	Norma y proceso inadecuados para la selección	CON 560; CON 540; SAN 110; SAN 100; SAN 210	Herramienta de ayuda; Fin público; Discriminador de sanciones; Identificación del sancionado; Fecha de resolución de la sanción
	Selección inadecuada	Selección inadecuada de beneficiarios de subsidios	CSU 120; CSU 130; PAG 110; DEV 110; INH 110; CON 630	Beneficiario; Discriminador de concesión de subvenciones; Discriminador de pagos; Discriminador de devoluciones; Discriminador de descalificaciones; Impacto de género
Ejecución	Transacciones no supervisadas	Transacciones que eluden los procedimientos de revisión normales o que no se controlan de otro modo	CON 502; CON 503; CON 504	Exención del Reglamento de la UE por categoría de ayuda; Objetivos de la exención; Regulación de exención por importe
	Pagos inconsistentes	El pago es excesivamente caro o no está relacionado con los objetivos del programa de subvenciones.	CSU 250; CSU 240; CSU 220; PRO 220; PRO 240: PRO 250; EJE 220; EJE 240; EJE 250	Ayuda equivalente a la concesión de subvenciones; Coste financiable de la actividad (subvención); Importe de la subvención; Importe de la subvención para el proyecto; Costes del proyecto; Ayuda equivalente (proyecto); Importe de la subvención para el organismo ejecutor por año; Coste del proyecto asignado al organismo ejecutor por año; Ayuda equivalente (ejecutor)
	Pagos redondeados	Un beneficiario de una subvención de reembolso que extrae fondos de la subvención utilizando números redondeados a la centena, millar o superior más cercana puede indicar que los fondos no se están extrayendo sobre una base de reembolso.	CSU 250; CSU 240; CSU 220	Ayuda equivalente a la concesión de subvenciones; Coste financiable de la actividad (subvención); Importe de la subvención

Fase 2	Nombre del indicador	Definición del indicador	Variables (código)	Variables (encabezado)
Seguimiento	Sanciones	Número elevado de infracciones sistemáticas por parte del destinatario	SAN 250, SAN 280; SAN 440; SAN 450	Multa por infracciones menores; Multa por infracciones graves; Publicidad de sanción; Plazo para publicitar la sanción

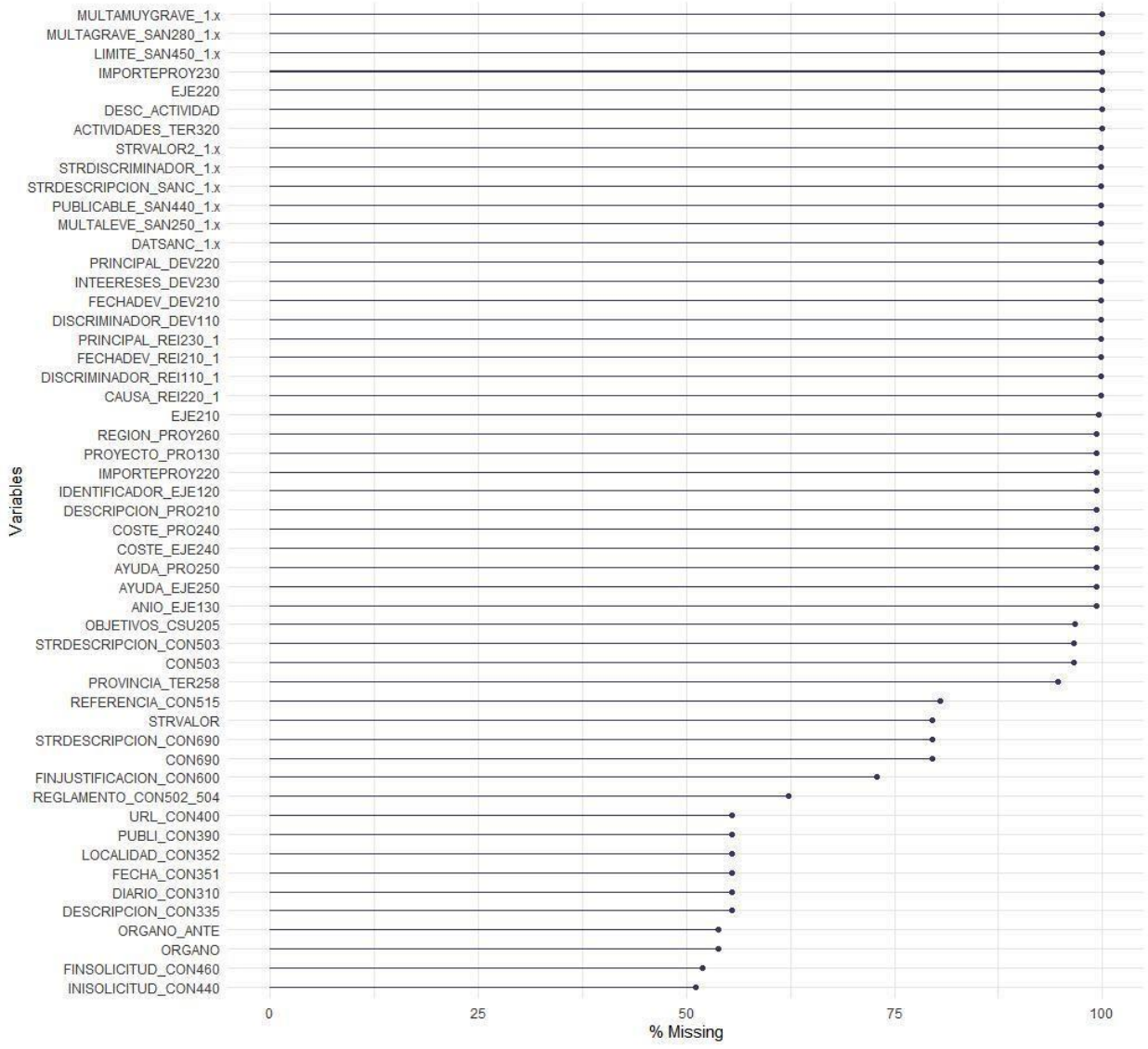
Fuente: Autor

Fusionar, limpiar y comprender las limitaciones de los datos

El primer paso del tratamiento de datos ha sido fusionar todos los conjuntos de datos facilitados por la administración española en el conjunto de datos principal que abarca todas las convocatorias y concesiones, BDNS_CONCESIONES. Para hacer eso, cada conjunto de datos se alineó con la misma unidad de análisis: la ID de la concesión. Cuando varias observaciones estaban relacionadas con la misma ID de concesión (por ejemplo, una concesión relacionada con varios sectores económicos), los datos se agregaban, por ejemplo, colocando cada observación duplicada en una columna aparte. Cuando las observaciones estaban relacionadas con una serie de ID de concesiones (por ejemplo, cuando el conjunto de datos contenía información sobre convocatorias), las características relevantes se copiaban en todas las concesiones relacionadas con esa observación de nivel superior. El conjunto de datos combinado, pero sin limpiar, contiene 1 792 546 concesiones y 152 variables. La lista completa de variables incluidas se puede encontrar en el Anexo B.

El siguiente paso del tratamiento de datos fue la limpieza de datos. Esto implicó eliminar variables con un índice alto de omisión o baja varianza. Estos problemas de datos afectan a un gran número de variables, como se muestra en la Figura 2.1. Se eliminaron todas las variables con un índice de omisión superior al 50 %, ya que habría introducido un mucho ruido en el análisis. La mayoría de estas variables con un índice alto de valores omitidos corresponden a sanciones y descripción del proyecto. Además, algunas de las variables mostraron una varianza muy baja, inferior a 0,3, lo que significa que contienen muy poca información relevante para el análisis posterior (es decir, en términos técnicos: su valor discriminante es bajo, ya que no varían lo suficiente entre observaciones sancionadas y no sancionadas). Por último, se eliminaron las variables de texto que no son directamente relevantes para la calificación de riesgo, como los descriptores de texto de las variables categóricas (por ejemplo, descripciones de sector económico) y las variables de texto libre con poca información relevante (por ejemplo, el título de la convocatoria).

Figura 2.1. Índices de valores omitidos



Fuente: Autor.

Como los métodos analíticos utilizados pueden ser sensibles a la información omitida, solo se conservaron aquellas observaciones que no tenían valores omitidos en todas las variables consideradas en el análisis, como las concesiones. Después de realizar todos estos pasos de tratamiento de datos, el conjunto de datos final utilizado en el análisis consta de 1 050 470 observaciones, concesiones y 60 variables para el periodo de 2018 a 2020 (inclusive).

Usar datos existentes para crear nuevos indicadores

Si bien la mayoría de los indicadores utilizados en el análisis se derivan directamente de los datos recibidos, algunos indicadores también se calcularon combinando otras variables. El primer grupo de estos indicadores calculados se refiere al importe y número de concesiones recibidas por el mismo beneficiario. El segundo grupo está formado por variables relacionadas con la ubicación: el nivel territorial del concedente y el beneficiario: nacional, regional o local. Además, se creó una variable binaria para identificar si la ejecución del proyecto se ubicó en el mismo lugar que el tercero. En tercer lugar, se calculó

un indicador que captura el mes de concesión de la subvención que puede indicar la periodicidad del gasto y los riesgos correspondientes. Por último, el sector económico del beneficiario se agrupó para recoger solo el nivel más alto de la clasificación NACE (Sección, categorías de 1 dígito). Véase en el Anexo A una tabla que describe todas las variables de estos cálculos de indicadores adicionales, y los pasos de tratamiento de datos detallados anteriormente. Esta es la lista final de variables utilizadas para el modelado de riesgos.

Definición de la variable dependiente en función del estado sancionador

La principal variable dependiente utilizada para el análisis es una variable binaria que indica si el beneficiario que recibe la concesión ha sido sancionado o no; con la sanción interpretada como una indicación fiable de fraude en una concesión. La variable pasa al valor «1» si el beneficiario ha sido sancionado por la concesión correspondiente, así como por todas las concesiones anteriores recibidas por el mismo, por haberse producido prácticas fraudulentas con anterioridad a la fecha de sanción. En caso de que el tercero no fuera sancionado, la concesión correspondiente obtiene el valor «0» en la variable ficticia. Las clases en la variable de sanción están muy desequilibradas: muestra 1031 casos de sanciones frente a 1 049 439 casos de ausencia de sanciones.

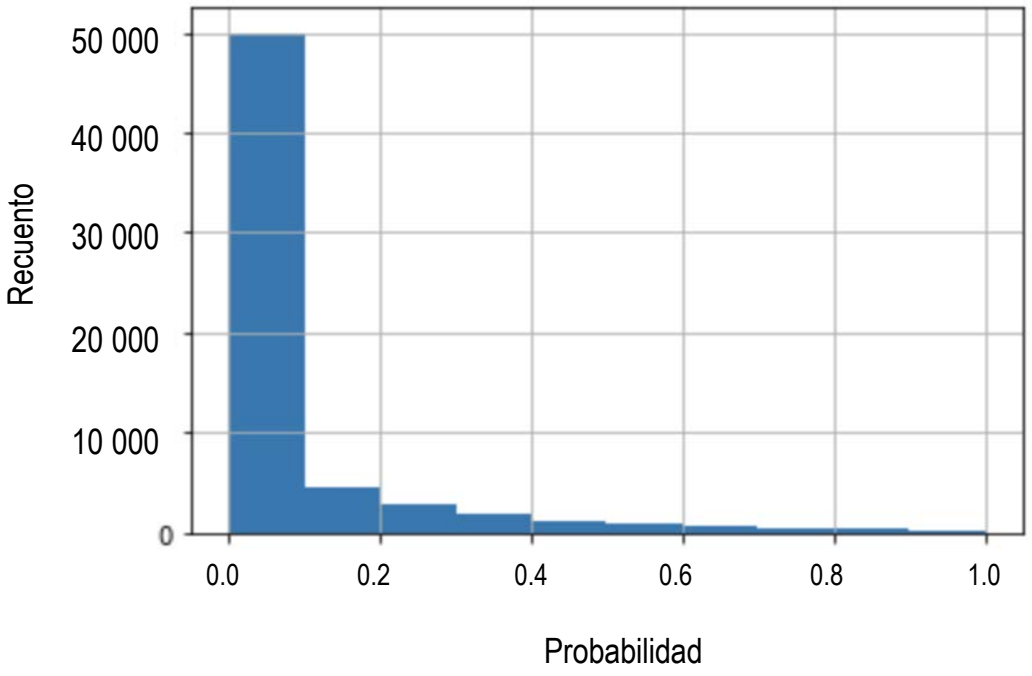
Para que el algoritmo de *random forest* se ejecute de manera eficiente, se ha extraído una muestra aleatoria de 90 000 concesiones de la parte sin etiquetar del conjunto de datos. Por tanto, el conjunto de datos de entrenamiento utilizado en el análisis hace uso de la muestra inicial de 91 031 concesiones, que consta de 1031 concesiones positivas (sanciones conocidas) y 90 000 concesiones sin etiquetar (estado de fraude poco claro).

Asignar el estado de sanción a las concesiones sin etiquetar

Para asignar etiquetas positivas y negativas a las observaciones sin etiquetar, se utilizó la metodología de aprendizaje positivo sin etiquetar. Este método comienza creando un subconjunto de entrenamiento de los datos que consta de todos los casos positivos, y una muestra aleatoria de casos sin etiquetar. Sobre esta muestra, el *bagging* PU construye un clasificador que asigna la probabilidad de sanción a cada concesión, a partir del cual es posible asignar la etiqueta positiva y negativa (probabilidad de sanción >50 % → etiqueta positiva). Estos pasos se repiten 1000 veces para construir un clasificador fiable que identifique los casos probables negativos y probables positivos en la muestra sin etiquetar (se debe tener en cuenta que la probabilidad de sanción media pronosticada en todos los modelos se convertirá en la calificación final pronosticada).

Como resultado de la ejecución de estos algoritmos, todos los casos sin etiquetar han recibido una probabilidad de sanción y, por tanto, una etiqueta de sanción probable (positiva frente a negativa). Para el conjunto de datos de entrenamiento, la Figura 2.2 presenta la distribución de las probabilidades de sanción (es decir, fraude). Esto pone de manifiesto que la mayoría de concesiones se consideran de bajo a muy bajo riesgo, y solo unas cuantas reciben una calificación de alto riesgo. En otras palabras, la mayoría de concesiones pueden clasificarse como no sancionadas, mientras que muy pocas concesiones reciben la etiqueta de sancionadas. En comparación con la muestra inicial positiva sin etiquetar, el número de casos probables positivos (sancionados) aumentó a 4430 con 86 601 identificados como probablemente negativos (no sancionados).

Figura 2.2. Clasificador de *insaculación* PU: predicción de probabilidad de sanción en la muestra inicial



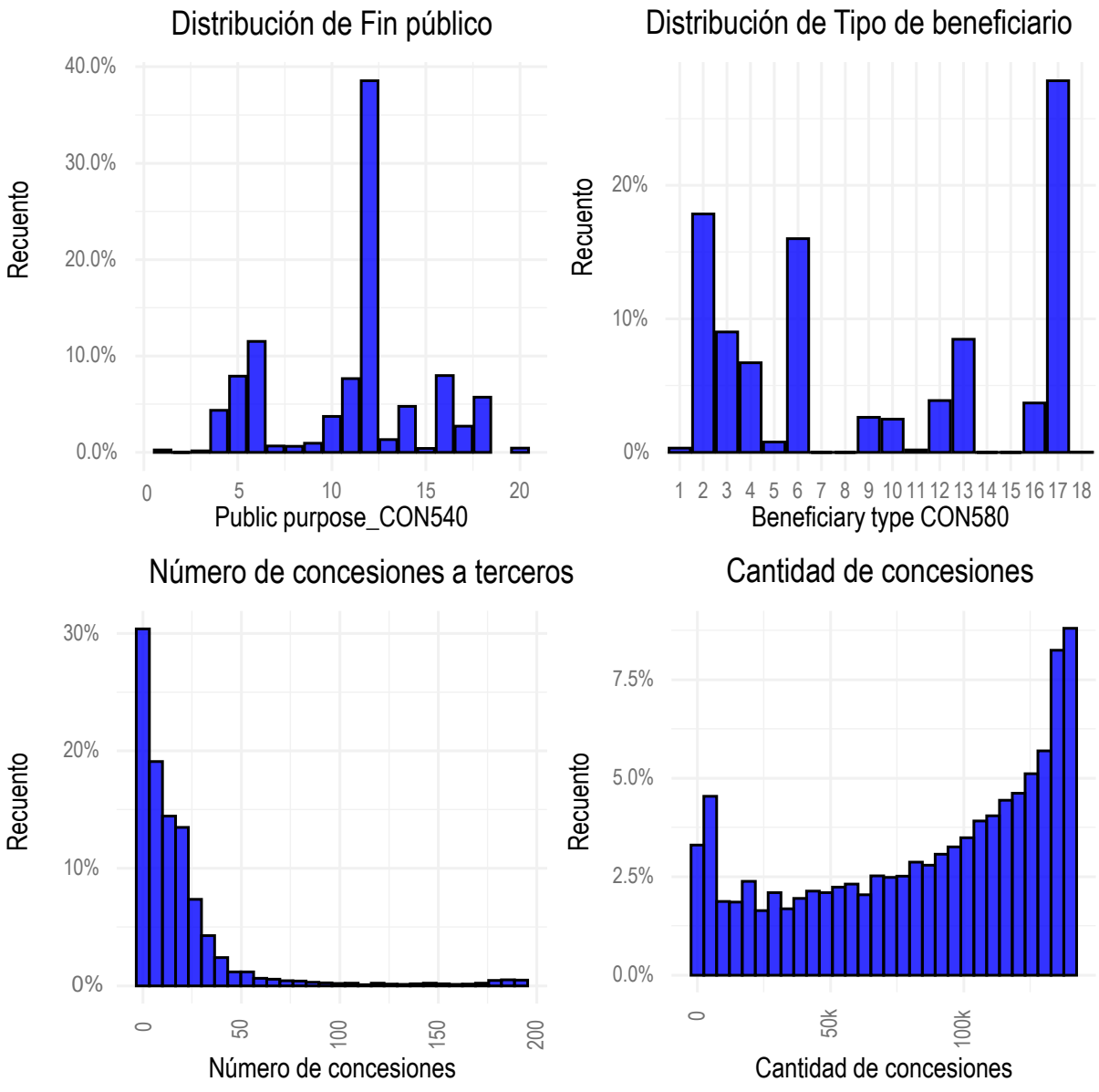
Fuente: Autor

Identificar las variables de mayor más impacto

Una vez que el conjunto de datos de concesiones positivos y sin etiquetar se vuelve a etiquetar y solo quedan casos positivos y negativos en el conjunto de datos siguiendo los métodos anteriores, se ejecuta un nuevo modelo de *random forest* y se comprueba su precisión. Esto significa que el conjunto de datos reetiquetado de 91 031 concesiones se dividió en una muestra de entrenamiento (70 %) y una muestra de verificación (30 %). El algoritmo *Random Forest* se entrenará en la primera y probará su precisión en la otra muestra que no ha «visto». El modelo óptimo consta de 1000 árboles y utiliza 106 variables en cada ejecución.

Este modelo óptimo de *random forest* ha identificado las variables más importantes para predecir la probabilidad de sanciones. A efectos de modelización, cada variable categórica se transformó en un conjunto de variables binarias, de modo que correspondan a una sola categoría de la variable categórica. Las variables numéricas se utilizaron tal cual, sin transformación. Las variables de mayor impacto en el modelo óptimo de *Random Forest* son Public_purpose_CON540, Nawards_TER_110, Amount_awards_TER110 y Third_party_legal_Spain_TER280. Sus distribuciones se presentan en la Figura 2.3.

Figura 2.3. Distribuciones de las variables de mayor impacto



Fuente: Autor

Estas distribuciones muestran que muchas de las variables más importantes tienen distribuciones significativamente desiguales. Por ejemplo, el número de subvenciones cae estrepitosamente por debajo de 50 y muy pocos beneficiarios tienen más de 50 subvenciones concedidas. De manera similar, la variable de fin público tiene un pequeño número de categorías predominantes, como 12 (agricultura). Además, el número de concesiones recibidas por el mismo beneficiario no está correlacionado con el valor total de las concesiones, lo que significa que la cantidad media de concesiones distribuidas es relativamente baja y algunas concesiones tienen un valor muy alto. La siguiente sección va un paso más allá y analiza los impactos de estas variables en la probabilidad de sanción (fraude).

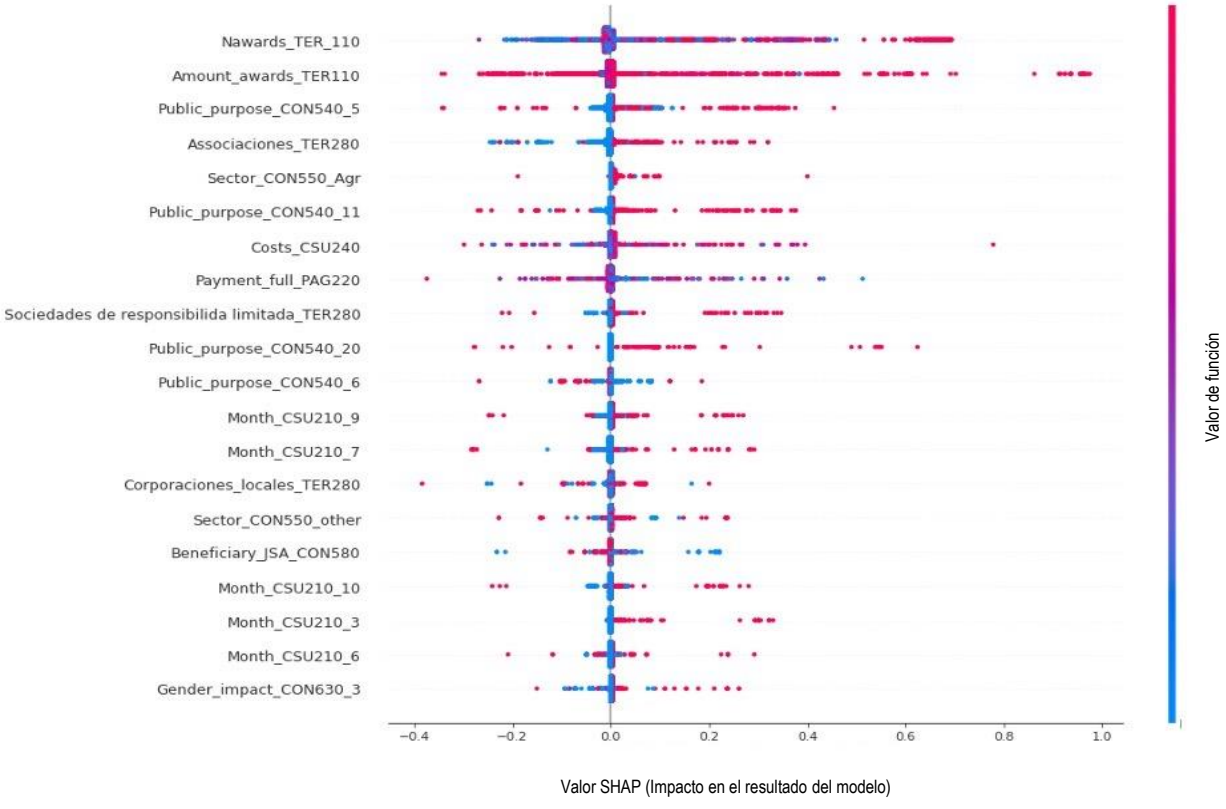
Probar el modelo en un conjunto de datos ciego

El modelo óptimo entrenado en el conjunto de datos de entrenamiento se ha verificado sobre datos no usados anteriormente, el conjunto de prueba (30 % de la muestra). Sobre este conjunto de datos de prueba, el modelo óptimo de *random forest* alcanzó: [5]

- exactitud = 95 % (la exactitud es el número de etiquetas predichas correctamente de todas las predicciones realizadas), y
- repetición = 93 % (es el número de etiquetas que el clasificador identificó correctamente dividido por el número total de observaciones con la misma etiqueta).

Estos resultados nos llevan a la conclusión de que el modelo es de gran calidad. Después de determinar la calidad general del modelo, la atención se ha centrado en el impacto de los predictores individuales en la probabilidad de sanción (fraude). Se debe tener en cuenta que los modelos de *Random Forest* capturan una gama de efectos interactivos y no lineales, por lo que interpretar las relaciones entre los predictores y el resultado es un asunto polifacético y complejo. Para mostrar el impacto de cada predictor de impacto en el resultado del modelo se siguió la literatura más reciente sobre aprendizaje automático, se calcularon los valores de explicaciones aditivas de Shapley (SHAP) (Lundberg and Lee, 2017[6]) y se han representado gráficamente. Los valores de SHAP ayudan a identificar la contribución individual de cada característica al modelo y su importancia para la predicción. El gráfico de Shapley en la Figura 2.4 muestra la probabilidad de sanciones (es decir, probable fraude) en función de los diferentes valores de cada predictor de impacto.

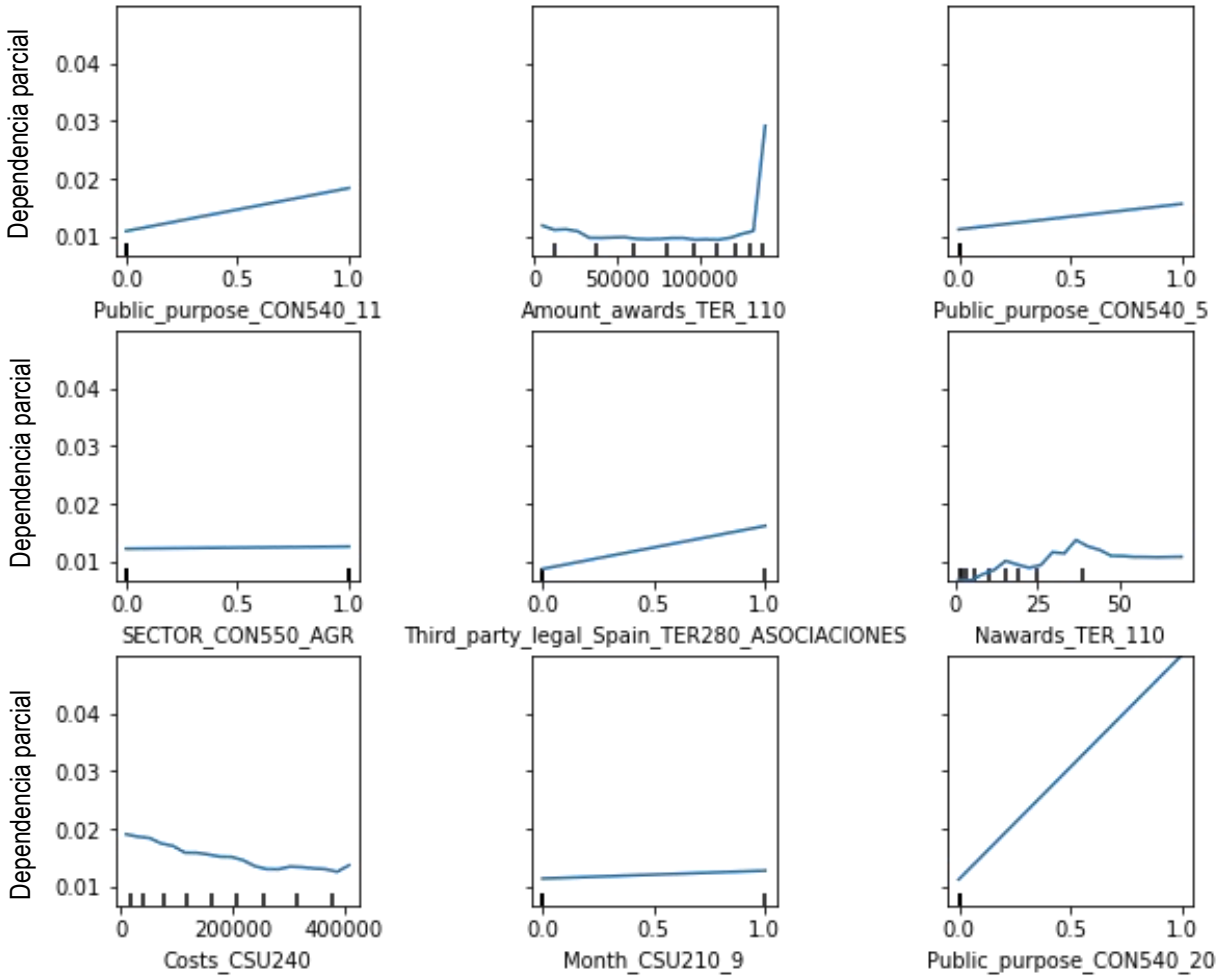
Figura 2.4. Valores de SHAP: Importancia variable y dirección del efecto



Fuente: Autor

La Figura 2.4 destaca que el impacto positivo más significativo en la probabilidad de sanciones lo proporciona el número de concesiones, así como el valor total de las concesiones recibida por el mismo beneficiario. Respecto al resto de predictores, la probabilidad de sanciones se correlaciona positivamente con la asociación y las sociedades limitadas como forma jurídica de beneficiarios, así como con el sector agrario en el sector económico. Los importes de la subvención están asociados negativamente a la probabilidad de ser sancionados, lo que significa que los importes más altos de los proyectos no se correlacionan con riesgos más elevados. Por el contrario, los fines públicos de la concesión, como la cultura (11), los servicios sociales (5), la cooperación internacional para el desarrollo y la cultura (20) y el fomento de empleo (6) están relacionados con una mayor probabilidad de sanciones. Además, las subvenciones concedidas en septiembre y julio se asocian a mayor probabilidad de sanción, con una tendencia similar en octubre, marzo y junio. Se muestran visualizaciones más detalladas de la influencia de las variables importantes en la probabilidad de sanciones (fraude) en la Figura 2.5.

Figura 2.5. Gráficos de dependencia parcial que representan el impacto de las variables seleccionadas en la probabilidad de fraude



Fuente: Autor

Finalización de la lista de indicadores para el modelo de riesgo

Para completar la descripción del modelo de evaluación de riesgos, se incluye el listado final de 29 indicadores válidos utilizados por el modelo según seis grupos (Tabla 2.3), haciendo referencia a las fases en las que podría ocurrir el posible fraude o las características de las organizaciones participantes: Fases de competición, selección, ejecución y seguimiento; organismo que concede la subvención y organización destinataria (beneficiario).

Tabla 2.3. Lista final de indicadores

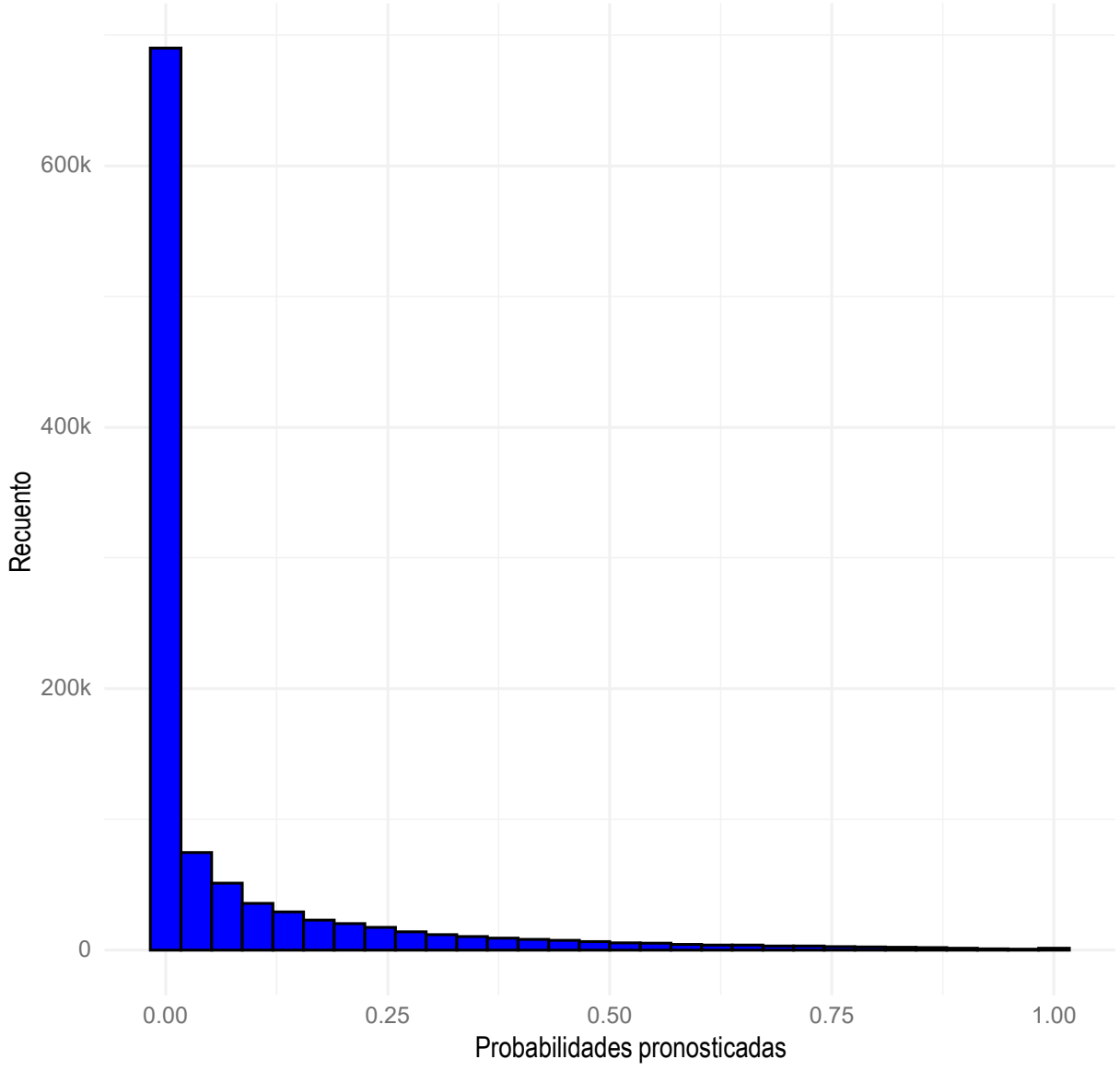
Fase	Variable	Descripción de la variable	Fraudes
Fase de competición	CON420, CON490, CON620	Admisión abierta, Condición de ayuda estatal, Convocatoria pública	La ausencia de admisión abierta o convocatoria pública conduce a un proceso de seguimiento menos transparente y, por tanto, existe mayor predisposición a actividades fraudulentas.
Fase de selección	CON540, CON580, CON610, CON630, SECTOR_CON550_AGR...EXTRATER, Month_CSU210	Fin público, Tipo de beneficiario, Subvención nominativa, Impacto de género, Sector de la economía, Mes de concesión	El tipo y la fecha de la convocatoria, el sector de la economía y el tipo de beneficiario podrían estar correlacionados con determinadas prácticas fraudulentas
Subvención/ejecución	CSU240, CSU220, CSU250, PAG220, PAG230, CON560, LOCAL_IMPL	Subvención nominativa, Importes, Concesión de subvención, Ayuda, Importe total pagado, Retención de impuestos, Instrumento de ayuda, Implantación local	Las subvenciones de importe elevado podrían ser potencialmente más propensas a actividades fraudulentas. Si la implantación se lleva a cabo en el mismo lugar que el concedente, podría ser una señal de un esquema de corrupción.
Organismo concedente	NATIONAL_CSU260, REGIONAL_CSU260, MUNICIAPAL_CSU260	Nivel de concesión de subvenciones	Las capacidades administrativas en determinadas regiones podrían ser insuficientes para un seguimiento eficaz de la convocatoria.
Organización destinataria	TER100, TER250, TER280, TER290, NATIONAL_TER310, REGIONAL_TER310, MUNICIAPAL_TER310, Amount_awards_TER110, Naward_TER110	País de terceros, Ubicación de terceros, Naturaleza jurídica de terceros, Tipo de terceros, Nivel de terceros, Número de concesiones, Cantidad de concesiones	La estructura y el tipo de organización de terceros, así como la ubicación, podrían estar correlacionadas con actividades fraudulentas. Las partes que reciben más concesiones de mayor tamaño podrían ser potencialmente más fraudulentas que otras.
Seguimiento	SAN_dum	Concesiones sancionadas	Captura la actividad fraudulenta del tercero

Fuente: Autor

Presentación de resultados y consideraciones para un desarrollo ulterior

La potencia de la metodología de evaluación de riesgos propuesta se muestra mejor utilizando el modelo óptimo final de *Random Forest* para asignar una calificación de riesgo de fraude a todas las concesiones de 2018 a 2020 con suficiente calidad de datos. Por tanto, la distribución final de la probabilidad de sanciones se presenta en la Figura 2.6 para las 1 050 470 concesiones observadas. En esta amplia muestra, el modelo predice que no habrá fraude (sanciones = 0) para 1 008 318 concesiones, mientras que predice fraude (sanciones = 1) para 42 152 concesiones utilizando el umbral del 50 % de probabilidad de sanciones para distinguir entre sanciones y no sanciones.

Figura 2.6. Distribución de probabilidades pronosticadas para todas las concesiones, nivel de concesión, 2018-2020

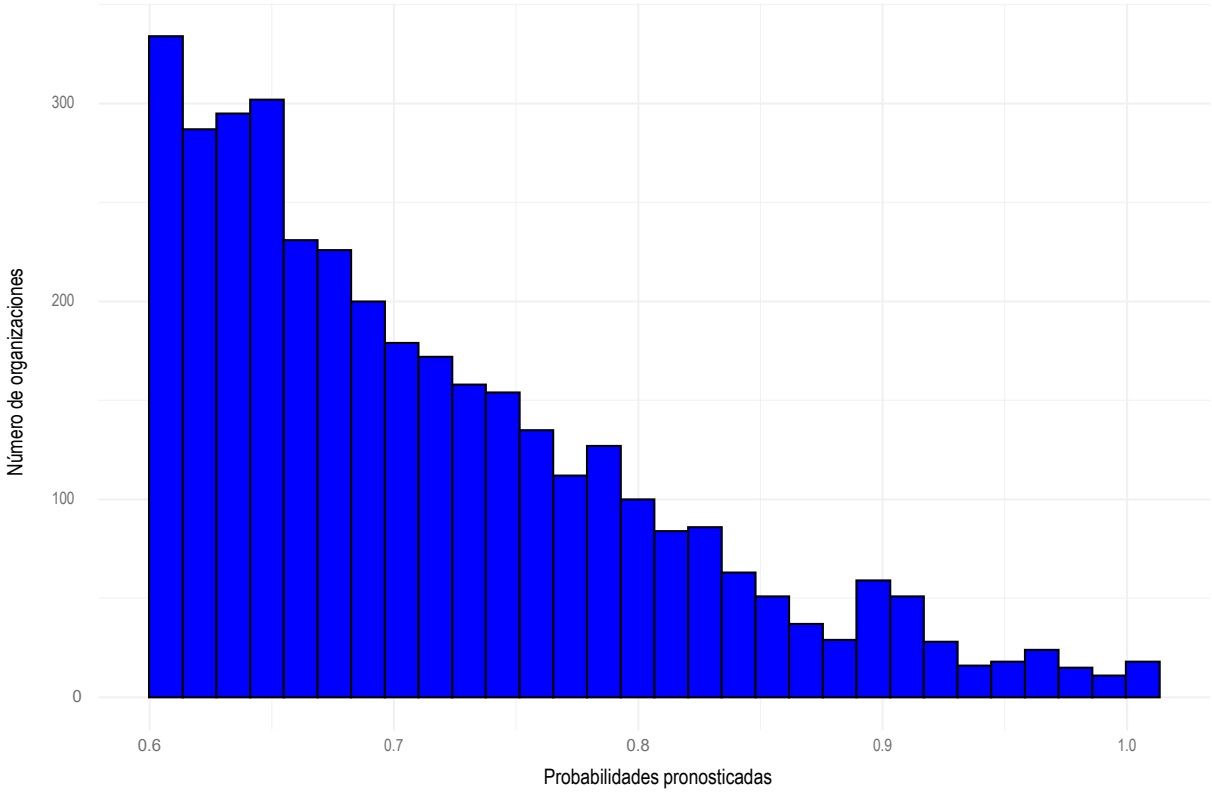


Fuente: Autor

Dado que los riesgos tienden a agruparse a nivel de las organizaciones y que las investigaciones suelen examinar todas las subvenciones recibidas por una organización, la exploración de las predicciones de las probabilidades de fraude a nivel de los beneficiarios añade valor al modelo. Para ofrecer una visión general de este nivel de agregación, mostramos la distribución de las predicciones de los riesgos de fraude por beneficiario con probabilidades de alto riesgo en la Figura 2.7. Ésta muestra que entre los beneficiarios de alto riesgo, las probabilidades de riesgo se distribuyen de forma desigual. La mayor parte de los beneficiarios de subvenciones de alto riesgo tienen características que indican una probabilidad de entre el 60% y el 70% de ser fraudulentos, y un grupo muy pequeño de organizaciones situadas en la cola derecha de la distribución tienen una probabilidad de ser fraudulentos de casi el 100% según el modelo. Estas organizaciones, las 10 primeras de las cuales se muestran en la Tabla 2.4, presentan el mayor riesgo y son las candidatas más adecuadas para un examen más profundo y una posible investigación

basada en el modelo predictivo. Además de éstas, las organizaciones a las que la IGAE decida seguir investigando dependerán de dónde fije su umbral de riesgo, y potencialmente de otros factores, como las implicaciones financieras (véase la sección « Combinar las puntuaciones de riesgo pronosticadas con la información financiera »)

Figura 2.7. Distribución media de probabilidades pronosticadas para organizaciones de alto riesgo, nivel de terceros, 2018-2020



Fuente: Autor

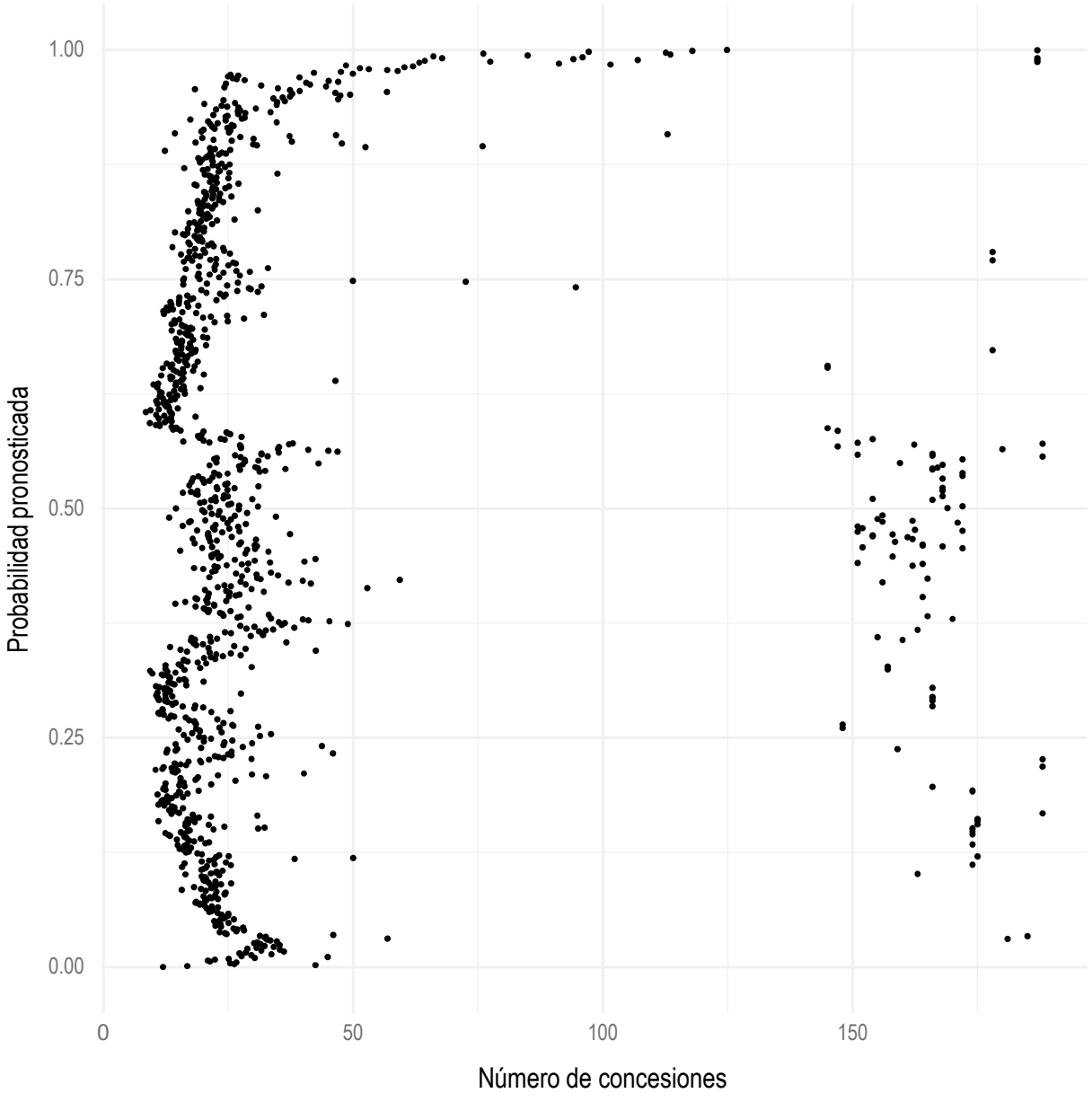
Tabla 2.4. Las 10 organizaciones principales por valor medio de concesiones

ID generada	Probabilidad pronosticada (media por tercero)
22568	1
46462	1
60626	1
101336	1
102140	1
129947	1
144235	0,996
152526	0,988
159661	1
167691	1

Fuente: Autor

Como era de esperar, el modelo estima que la inmensa mayoría de las concesiones no presentan riesgos, pero algunos miles de concesiones se marcan como de riesgo, además de las 1031 concesiones observadas sancionadas. Teniendo en cuenta las variables más importantes del modelo, se analiza más de cerca la distribución de las probabilidades de fraude. Primero, la Figura 2.8 muestra la distribución del número de concesiones recibidas por un mismo beneficiario en relación con su probabilidad de sanción. Curiosamente, el modelo predice una alta probabilidad de sanción para entidades receptoras grandes y pequeñas. La mayoría de concesiones se ubican en el lado izquierdo del gráfico, con 0 a 50 concesiones por beneficiarios y probabilidades relativamente igualadas de ser sancionadas para este grupo de observaciones. A partir de 50 concesiones, la probabilidad aumenta a casi el 100 %, con una disminución a alrededor del 50 % cuando el número supera las 150 concesiones. Esto podría explicarse por la fiabilidad de beneficiarios: si se demuestra que estas organizaciones son fiables durante largo tiempo reciben más concesiones. Mientras que para las primeras 50 concesiones, se lleva a cabo un proceso de evaluación. También es concebible que después de un umbral determinado de 50 concesiones por beneficiario, las investigaciones se lleven a cabo con mayor frecuencia y, por tanto, es más probable que aparezcan concesiones potencialmente sancionadas.

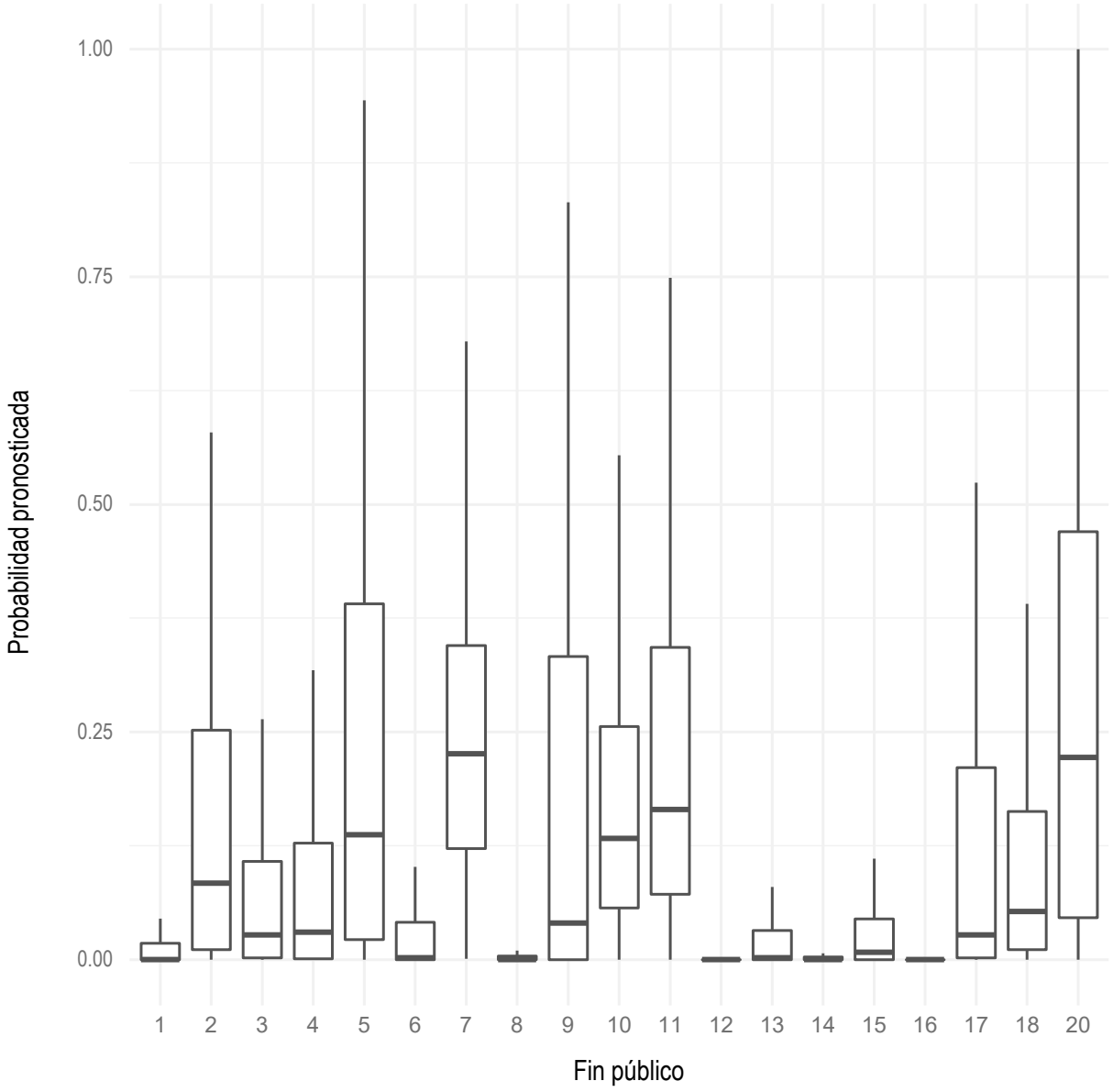
Figura 2.8. Distribución del número de concesiones por probabilidad de sanciones



Fuente: Autor

En segundo lugar, respecto a otra variable importante, el fin público de la convocatoria, se presenta su distribución de probabilidades en la Figura 2.9. Dos categorías muestran el mayor riesgo de sanciones: servicios sociales (5) y cooperación internacional para el desarrollo y la cultura (20). Es importante destacar que estas no son las categorías más frecuentes entre las concesiones; la más frecuente es la agricultura (12), que muestra el riesgo más bajo.

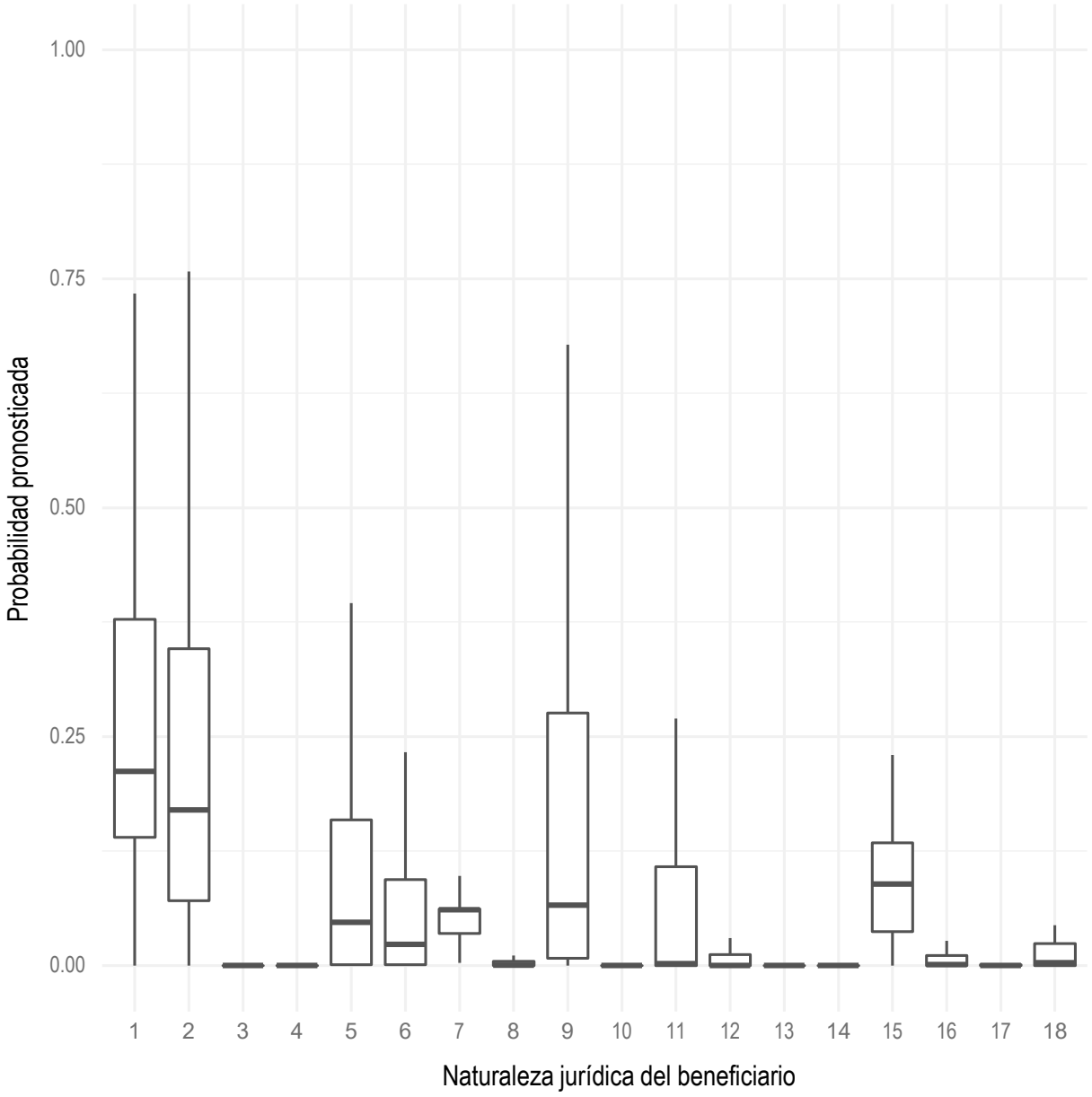
Figura 2.9. Distribución del fin público de la convocatoria sobre la probabilidad de sanciones



Nota: 1 - Justicia, 2- Defensa, 3 - Seguridad ciudadana e instituciones penitenciarias, 4- Otros beneficios económicos, 5- Servicios sociales y promoción social, 6- Fomento de empleo, 7- Desempleo, 8- Acceso a la vivienda, 9- Salud, 10- Educación, 11- Cultura, 12- Agricultura, Pesca y Alimentación, 13- Industria y Energía, 14- Comercio, Turismo y Pymes, 15- Subsidios para transporte, 16- Infraestructura, 17- Investigación, Desarrollo e Innovación, 18- Otras acciones económicas, 20 - Cooperación internacional para el desarrollo y la cultura
Fuente: Autor

En tercer lugar, la naturaleza jurídica del beneficiario es otra variable importante identificada por el modelo (Figura 2.10). La segunda categoría - asociaciones - mostró un impacto positivo significativo en la probabilidad de sanciones en el modelo presentado. Otros dos tipos de beneficiarios también son propensos a mayores riesgos: los órganos de la administración estatal y las comunidades autónomas (1) y los organismos públicos (9). Si bien la asociación es también la categoría más frecuente para esta variable, las categorías 1 y 9 son las menos frecuentes, pero muestran una gran probabilidad de ser sancionadas.

Figura 2.10. Distribución de la naturaleza jurídica de terceros sobre la probabilidad de sanciones

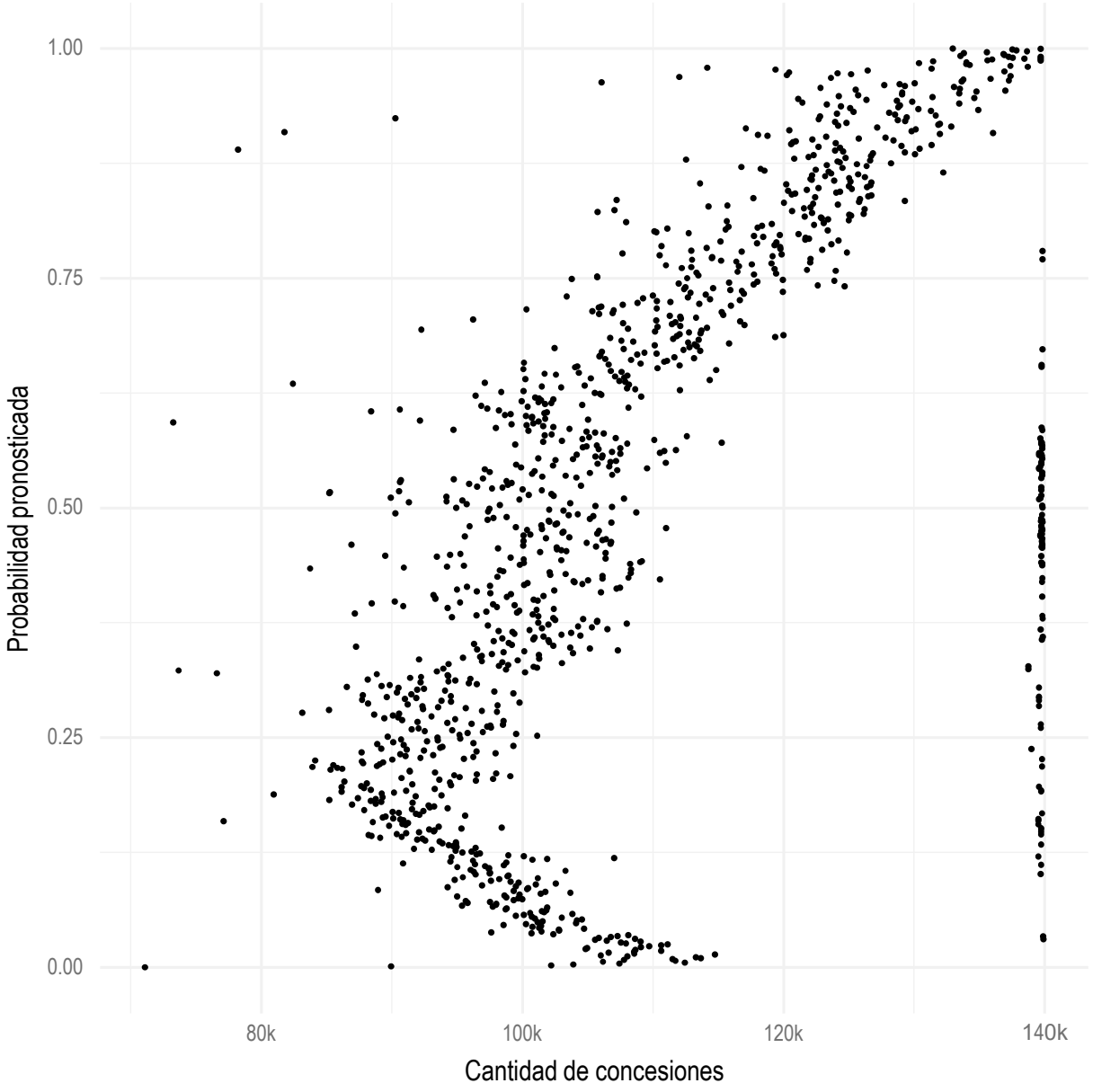


Nota: 1 - Órganos de la administración estatal y comunidades autónomas, 2 - Asociaciones, 3 - Comunidades de propiedad, herencias y otras entidades sin personalidad jurídica, 4 - Comunidades de propietarios en régimen de propiedad horizontal, 5 - Instituciones religiosas, 6 - Administración local, 7 - Entidad extranjera, 8 - Establecimiento permanente de entidad no residente en territorio español, 9 - Organismos públicos, 10 - Otros tipos, 11 - Persona jurídica con identificación no generada por autoridades españolas (AEAT o Policía), 12 - Sociedades anónimas, 13 - Organizaciones civiles, 14 - Organizaciones colectivas, 15 - Sociedades comandadas, 16 - Sociedades cooperativas, 17 - Sociedades limitadas, 18 - Uniones temporales de empresas
Fuente: Autor

Por último, también se ha encontrado que el importe total de las concesiones recibidas por el beneficiario tiene un impacto significativo en la probabilidad de sanciones (Figura 2.11). Hay un crecimiento sostenido de las probabilidades de sanción a partir de 90 000 €. Además, existe una divergencia en las probabilidades previstas entre 85 000 € y 110 000 €, lo que demuestra que hasta los 110 000 €, no todas las concesiones presentan riesgos. Por último, para el importe total máximo de las concesiones

observadas (140 000 €), la probabilidad de sanciones se distribuye uniformemente entre 0,05 y 0,76. Esto es muy similar a lo que se observó en la distribución del número de concesiones: el número más alto se asocia a una distribución uniforme de los riesgos.

Figura 2.11. Distribución del importe total de las concesiones recibidas por el mismo tercero sobre la probabilidad de sanciones



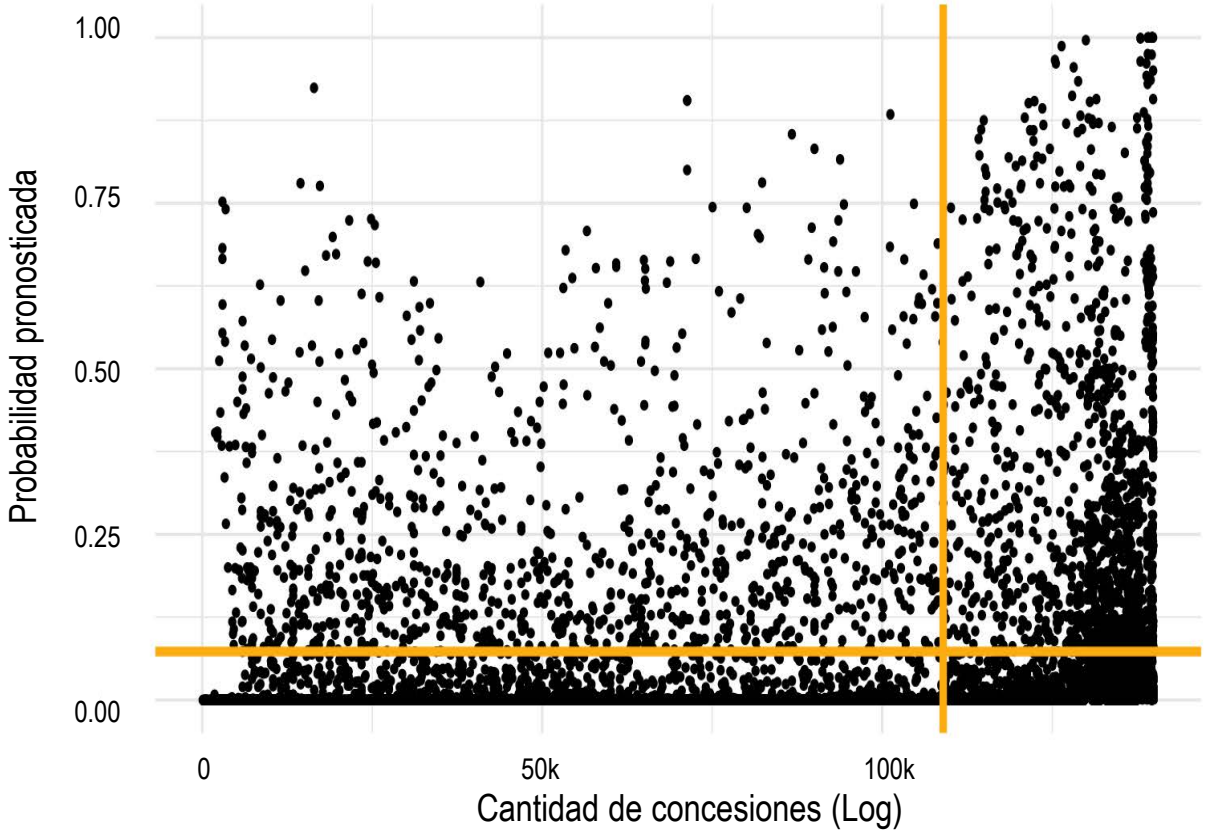
Fuente: Autor

Combinar las puntuaciones de riesgo pronosticadas con la información financiera

Los riesgos de fraude representan la variable clave de interés para IGAE y, por tanto, sirven como la principal variable dependiente para el modelo descrito hasta ahora. Sin embargo, solo representan una de las dimensiones claves según las cuales se pueden seleccionar los objetivos de la investigación. Una

segunda dimensión clave que podría tenerse en cuenta es el valor total de la subvención como una indicación del posible impacto económico del fraude para el gobierno español. La combinación de las puntuaciones de riesgo de fraude estimadas con el valor total de la concesión permite a quienes toman las decisiones y a los investigadores tener en cuenta, simultáneamente, la prevalencia de los riesgos y sus probables consecuencias económicas. (Fazekas, M., Ugale, G, & Zhao, A., 2019[7]). El enfoque más sencillo para observar estas 2 dimensiones simultáneamente es dibujar un diagrama de dispersión con estas 2 variables destacando también sus valores medios (Figura 2.12). El cuadrante superior derecho incluye aquellas concesiones que no solo tienen un alto riesgo, sino que también tienen valores altos de concesiones. Este es el grupo de mayor interés para las futuras investigaciones de la IGAE, ya que es más probable que incluyan subvenciones fraudulentas con importantes consecuencias económicas.

Figura 2.12. Distribución de concesiones según la puntuación de riesgo pronosticado y el valor total de la concesión



Fuente: Autor

Establecer un conjunto de datos preparado para detectar riesgos de fraude en el futuro

Para seguir mejorando el marco de evaluación de riesgos de fraude basado en datos de la IGAE, se pueden implantar una serie de reformas a corto y medio plazo que mejoren la calidad y el alcance de los datos subyacentes a los modelos de riesgo. El desarrollo de un modelo de riesgo por parte de la OCDE ya ha ayudado a la IGAE a avanzar en el tratamiento de algunos de estos problemas, y el conjunto de datos resultante del trabajo de la OCDE puede ser un punto de partida para la IGAE. No obstante, los

conjuntos de datos no son estáticos y pueden aparecer disponibles nuevas fuentes de datos. Por tanto, estos puntos son relevantes fuera del contexto de este proyecto.

Primero, los datos existentes pueden combinarse mejor y más rápido en un solo conjunto de datos preparado para modelar el riesgo de fraude. Actualmente, casi todos los conjuntos de datos abarcan distintas unidades de análisis como concesiones, convocatorias u organizaciones. Para que la IGAE los fusione, cada conjunto de datos debe alinearse al mismo nivel con ID únicas para evitar la multiplicación redundante de observaciones en el conjunto de datos fusionado. Durante el tratamiento de datos para este informe, estos se han transformado de formato largo a ancho cuando ha sido necesario. Sin embargo, este enfoque adolece de un gran inconveniente, que es un índice alto de omisión de ID sin múltiples observaciones por ID única. Para resolver este problema, se necesita la agregación, especialmente para las variables factoriales que no pueden calcularse como medias o medianas.

En segundo lugar, es fundamental reducir las tasas de omisión en todas las variables recopiladas por la IGAE. Como se trató en el Capítulo 1, definir estándares de calidad de datos y aplicarlos, en colaboración con los propietarios de datos, aseguraría que no haya variables con tasas de omisión elevadas, como 40 %-50 %. En tercer lugar, algunos conjuntos de datos (por ejemplo, sobre proyectos) consisten en una cantidad muy pequeña de observaciones, lo que impide su análisis junto con el conjunto de datos principal (es decir, cuando se combinan, dan como resultado un índice de omisiones alto). Del mismo modo, los datos sobre beneficiarios son muy limitados y necesitan mejoras adicionales.

Ampliar el uso de indicadores por parte de la IGAE a lo largo del ciclo de subvenciones.

Como se señaló, la lista final de indicadores incluye 29 variables. Estas variables son en su mayoría categóricas, aunque existen algunas variables numéricas, como importes y pagos. La mayoría de las variables analizadas son descriptivas debido a los datos disponibles. Hay datos limitados que pueden proporcionar información sobre los comportamientos de organizaciones y personas, como los conflictos de intereses entre los actores que reciben o se benefician de las subvenciones. Esta es una de las mayores lagunas en los datos actuales disponibles para la IGAE y es uno de los factores más restrictivos en su análisis de riesgo, independientemente de la metodología utilizada. La Tabla 2.5 muestra indicadores de comportamiento adicionales que podrían usarse para la evaluación de riesgos de fraude que abarcan el ciclo de la subvención y podrían ayudar a refinar el modelo de riesgo de la IGAE.

Tabla 2.5. Indicadores de comportamiento para evaluar los riesgos de fraude en cada fase del ciclo de subvenciones

Grupo de indicadores	Nombre del indicador	Definición del indicador
Fase de competición	falta de competición	Solo un solicitante para una convocatoria de subvenciones
Fase de selección	concentración de importes	Cantidad y valor excesivos de pagos a un solo proveedor
	Influencia política	Solo a los beneficiarios vinculados al gobierno se les conceden sus solicitudes
Fase de ejecución de la subvención	sobrefinanciación	Salario u otra compensación por servicios personales que exceden los importes aprobados por la agencia o son más altos que la compensación por otros servicios comparables que no están financiados por subvenciones.
	grandes pagos anticipados	Un beneficiario que extrae todos o la mayoría de los fondos de la subvención poco después de la concesión de la subvención puede ser característico de un riesgo más alto, a menos que el programa de subvenciones permita esta práctica.
	Modificación de plazos	Solicitud del contratista para modificar los plazos y las condiciones del contrato.
	Operación grande	Una sola operación representa más de la mitad de los costes totales del proyecto.
	Gastos atrasados	Gastos fuera del período permitido del proyecto
	Operaciones inusuales	Las operaciones cuestionables o inusuales inmediatamente anteriores al final

Grupo de indicadores	Nombre del indicador	Definición del indicador
		del período de concesión de una subvención pueden indicar que los defraudadores esperaron hasta el final del proyecto para retirar los fondos de la subvención para cubrir los costes no permitidos.
Organización destinataria	empresa nueva	Beneficiario final constituido inmediatamente antes de la solicitud de la subvención
	doble financiación	Prueba de que los beneficiarios están financiando proyectos de subvenciones con más de una subvención
	Viabilidad económica	Un destinatario que tiene una viabilidad financiera cuestionable, como un alto porcentaje de activos financiados con deuda o una liquidez insuficiente.
Contratistas y asesores	adquisición no competitiva	Destinatarios que gastan fondos en compras no aprobadas o subcontrataciones de adquisición de fuente única no aprobados o sin licitación
	subcontrataciones de asesores	El uso de asesores genéricos, no específicos o confusos
	documentación insuficiente	Justificación y documentación insuficientes para pagos realizados a contratistas/asesores, como horas trabajadas y actividades
Seguimiento y auditorías	consultas de auditoría	Varias consultas de las fuerzas del orden o las oficinas de auditoría que no pueden responderse
	no cooperación con auditores	El personal receptor que no coopera con las actividades de seguimiento o es agresivo con los auditores o gestores de subvenciones

Fuente: Autor

Existe una variedad de fuentes y ejemplos que pueden ayudar a la IGAE a mejorar sus indicadores de riesgo. En la Unión Europea, la Oficina Europea de Lucha contra el Fraude (OLAF) creó en 2011 un Compendio de casos anonimizados, que todavía tiene relevancia en la actualidad. El Compendio enumera los resultados de las investigaciones de la OLAF e incluye información sobre fraudes financieros. Se pueden identificar dos fases de alto riesgo de comportamiento fraudulento potencial: la fase de selección y la fase de ejecución. Durante la fase de selección, la OLAF fomentó que se inspeccionaran de cerca declaraciones justificativas y documentación oficial, así como a asegurarse de que el beneficiario final no se haya constituido o creado inmediatamente antes de la publicación de la subvención. Durante la fase de ejecución, la OLAF sugirió tener en cuenta las dificultades financieras del contratista, la presencia de una única operación grande que cubra casi la mitad de todos los costes del proyecto, así como el uso de la subvención para otros fines (European Anti-Fraud Office (OLAF), 2011[8]). El Compendio ilustra la realidad de que muchos fraudes son simplemente versiones recicladas de sistemas similares. De hecho, en su *32º Informe anual sobre la protección de los intereses financieros de la Unión Europea - Lucha contra el fraude -2020*, la Comisión Europea señaló que, entre las irregularidades fraudulentas relacionadas con la infraestructura sanitaria y la pandemia de COVID-19, los problemas más frecuentemente detectados se referían a la documentación de apoyo (European Commission, 2020[9]). El Recuadro 2.2 proporciona información adicional a partir de la experiencia del Comité de Fraude en Subvenciones del Grupo de Trabajo de Lucha contra el Fraude Financiero, que se creó para abordar el fraude a raíz de la crisis económica de 2008.

Recuadro 2.2. El Comité de Fraude en Subvenciones del Grupo de Trabajo de Lucha contra el Fraude Financiero de EE. UU.

En Estados Unidos, el Comité de Fraude en Subvenciones del Grupo de Trabajo de Lucha contra el Fraude Financiero identificó varias áreas clave para monitorizar e identificar actividades fraudulentas:

- la estructura de la organización destinataria y el programa de subvenciones
- solicitudes de pago o retiradas de fondos ordinarios
- seguimiento de informes y actividades

- actividades a nivel de operaciones
- contratos y asesores.

Entre la primera categoría, el Comité de Fraude en Subvenciones sugirió monitorizar el diseño del proyecto, así como la viabilidad económica del destinatario, el control interno, el personal de las organizaciones y los posibles conflictos de interés. En lo que respecta a las solicitudes de pago, se debe prestar atención al momento de la concesión de la subvención, así como a la documentación justificativa, al exceso de gastos y al redondeo de las cifras para la concesión de la subvención. Al realizar las actividades de seguimiento, la capacidad de respuesta y la cooperación del beneficiario son indicadores clave, así como la presencia de controles internos y el historial de auditorías de la empresa. Cuando se trata de actividades a nivel de operaciones, las operaciones excesivas, inusuales y no supervisadas podrían marcarse como riesgos potenciales y también como doble financiación (más de una subvención que cubra el mismo proyecto). Por último, en lo que respecta a los contratos y asesores, el Comité de Fraude en Subvenciones sugiere examinar las operaciones de partes relacionadas, el gasto en asesores no específicos y los beneficiarios de subvenciones con deficiencias en sus sistemas de adquisiciones. En caso de monitorización de datos, el Comité de Fraude de Subvenciones (2012) identifica los siguientes riesgos de fraude:

- número e importe excesivos de pagos a un solo proveedor
- pagos a proveedores no aprobados
- operaciones que eluden los procedimientos de revisión normales o que, de otro modo, no los controla otra persona
- compras que parecen irracionales, teniendo en cuenta la naturaleza del programa de subvenciones
- gastos fuera del período permitido del proyecto
- pagos y operaciones recurrentes
- pagos emitidos a varios proveedores a la misma dirección postal

Nota: En 2018, el Grupo de Trabajo sobre Integridad del Mercado y Fraude al Consumidor sustituyó al Grupo de Trabajo de Ejecución de Fraude Financiero.
Fuente: (Financial Fraud Enforcement Task Force, 2012[10])

Invertir en la mejora continua del modelo de riesgo

Dado que la validez de este modelo basado en datos depende de las sanciones impuestas, si las sanciones no cumplieran con los esquemas de fraude relevantes o resultaran en una muestra sesgada de investigaciones, cualquier modelo de evaluación de riesgos también estaría sesgado. Por tanto, obtener una muestra verdaderamente aleatoria de investigaciones y sanciones es de vital importancia. Con este fin, la IGAE puede seleccionar un porcentaje de los casos investigados cada año utilizando sus técnicas tradicionales de muestreo o un método de selección basado en datos como el presentado anteriormente. El resto de los casos investigados puede elegirse mediante una selección aleatoria completa. Este enfoque lograría un equilibrio entre optimizar la utilidad de los recursos de investigación a través de una mejor focalización, mientras que también invertiría en futuras mejoras del modelo de evaluación de riesgos, al proporcionar una muestra de entrenamiento mejor. También le daría a la IGAE una mejor idea del desempeño del modelo. Como este esfuerzo presente ha sido una prueba conceptual, se pueden valorar pasos técnicos adicionales:

- Mejorar la calidad de las variables en el todo el conjunto de datos para poder incluir más indicadores en el modelo, y por tanto mejorar su calidad.

- Se debe tener en cuenta la naturaleza desequilibrada de las clases en la variable dependiente (positiva/negativa): utilizar técnicas de insaculación PU para evitar imprecisiones en el modelado.
- Repetir el ejercicio de modelización con regularidad a medida que se disponga de nuevos datos, incluidas las sanciones y las concesiones, para mantener actualizada la evaluación de riesgos.
- Los modelos analíticos no dibujan una imagen completa y pueden tener sesgos a medida que aprenden de las acciones de control pasadas (ver Capítulo 1). La IGAE podría complementar los modelos con métodos cualitativos y el criterio experto. Esto permite a los especialistas en fraude de la IGAE contribuir con su conocimiento especializado sobre sistemas de fraude, los últimos eventos y el contexto más amplio.

Si bien los modelos pueden ser vulnerables a los sesgos en sí mismos, también pueden ayudar a controlarlos. Concretamente, la selección de muestras basada en datos, incluidas las que utilizan el aprendizaje automático, no solo ayudaría a la IGAE a optimizar la eficacia de los recursos de control, sino que también ayudaría a corregir algunos sesgos en el conjunto de datos de entrenamiento. Por ejemplo, si se conocen los tipos de fraude que no están cubiertos por las investigaciones, sus características se pueden introducir manualmente en la base de datos para proporcionar información suficiente para que el algoritmo aprenda. Además, si la selección de investigaciones en el conjunto de sanciones enfatiza ciertas variables, como el importe de la subvención, submuestrear las subvenciones grandes y las subvenciones pequeñas puede contrarrestar la selección sesgada de los casos investigados.

Tener en cuenta los análisis de redes y utilizar un conjunto más amplio de metodologías

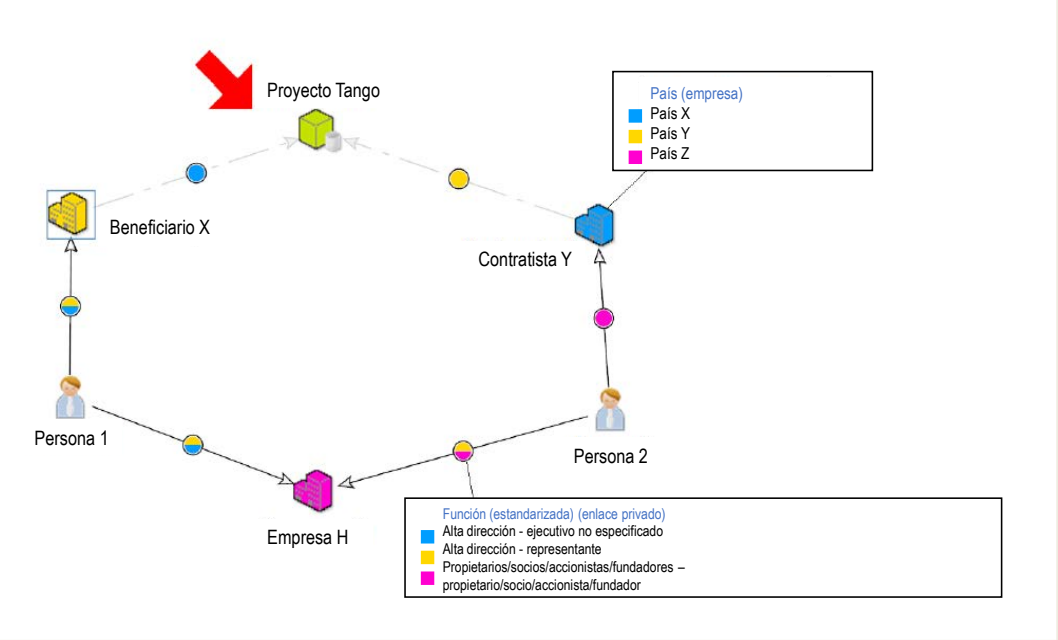
Las técnicas de ciencia de datos y redes se han utilizado cada vez más para estudiar los delitos económicos como la corrupción, el fraude, la colusión, la delincuencia organizada o la evasión fiscal, por mencionar algunas áreas importantes (Wachs, Fazekas and Kertész, 2020[11]). Explorar redes sin análisis avanzado ya promete grandes ventajas para la detección de fraudes, como el rastreo de posibles conflictos de intereses (ver el Recuadro 2.3).

Recuadro 2.3. Usar datos para investigar conflictos de interés

Cuando se conocen las personas que están detrás de las organizaciones públicas y privadas que participan en el proceso de concesión de la subvención, se puede descubrir una serie de posibles de relaciones con conflicto de interés. Si bien las investigaciones en profundidad pueden revelar dichas relaciones, la detección de riesgos se facilita en gran medida al hacer coincidir conjuntos de datos a gran escala que contienen: 1) todos los cargos públicos que desempeñan un papel importante en la preparación, evaluación, concesión y seguimiento de subvenciones; y 2) todas las personas físicas que tienen un papel importante en las empresas que presentan solicitudes, reciben y ejecutan subvenciones.

La recopilación, limpieza y vinculación de dichos conjuntos de datos y el mantenimiento de las conexiones subyacentes pueden generar costes importantes. Sin embargo, una vez que se dispone de un conjunto de datos de este tipo y una interfaz gráfica simple, como es el caso de la herramienta ARACHNE de la UE, se puede acelerar enormemente la selección e investigación de las relaciones de riesgo entre los concedentes y los beneficiarios de las mismas. Por ejemplo, es posible examinar de forma rápida y eficiente los proyectos, los beneficiarios y las personas que participan en la preparación de la convocatoria y la evaluación de las solicitudes.

Figura 2.13. Visualización de conflictos de intereses



Fuente: (Unión Europea, 2016[12])

El análisis a gran escala de redes de relaciones contractuales o personales puede revelar patrones ocultos que sirven como indicadores de riesgo por sí mismos o complementan otros indicadores de riesgo (Fazekas and Tóth, 2016[13]); (Fazekas and Wachs, 2020[14]). Por ejemplo, los indicadores de riesgo de licitación, como la ocurrencia de licitación única en la contratación pública, pueden superponerse a grupos de compradores y proveedores vinculados en la contratación pública para identificar grupos organizados de alto riesgo. La Figura 2.14 a continuación muestra una visualización de relaciones de compradores y proveedores en contratación pública en Hungría. Dichos diagramas proporcionan una instantánea visual de los datos que señalan las posibles relaciones de alto riesgo para una investigación ulterior. Por ejemplo, las líneas rojas resaltan un indicador de oferta única más alto que la media de ofertas únicas en esa relación. Además, hay un grupo de actores de alto riesgo de corrupción en la parte superior (es decir, relaciones densas de contratación que coinciden con índices altos de licitación única en esas relaciones).

Figura 2.14. Relaciones de compradores y proveedores en contratación pública, Hungría 2014

Fuente: (Wachs, Fazekas and Kertész, 2020[11])

La IGAE puede recopilar conjuntos de datos relevantes, como datos de propiedad de las empresas, y vincularlos a sus datos básicos de subvenciones, para hacer uso de dichas técnicas de análisis de redes. A medida que las personas se mueven entre los sectores público y privado y hay otras formas en que los beneficiarios pueden establecer conexiones con los organismos que conceden subvenciones, el rastreo de redes abiertas u ocultas ofrece una herramienta clave para mejorar la evaluación de riesgos de fraude en España. Como se ha tratado, esta es un área en la que la IGAE actualmente tiene lagunas en sus datos, por lo que el uso de análisis de redes también dependerá de la capacidad de la IGAE para abordar estas lagunas. Los acontecimientos recientes en España sugieren que ya se están realizando mejoras. Por ejemplo, en mayo de 2020, la IGAE y la Tesorería General de la Seguridad Social (TGSS) firmaron un convenio sobre transferencia de información, estableciendo condiciones más colaborativas para el control financiero de las subvenciones y ayudas públicas. El acuerdo estipula el acceso directo a las bases de datos de la TGSS para facilitar el trabajo de la IGAE en la detección de fraude e irregularidades (Ministry of the Presidency of Spain, 2021[15]). Avanzar en acuerdos similares con otras entidades públicas y privadas, concretamente para obtener datos empresariales y datos que reflejen indicadores de comportamiento, como se mencionó anteriormente, constituyen aportaciones fundamentales para fortalecer futuros modelos de riesgo.

Conclusión

Este capítulo ha presentado una prueba de concepto para que la IGAE mejore su enfoque de evaluación de los riesgos de fraude en las subvenciones públicas, basándose en las principales prácticas de análisis. El proceso de desarrollo del modelo de riesgo ha llevado a una serie de descubrimientos sobre la capacidad actual de análisis de la IGAE, así como de la gestión de datos y la garantía de calidad de los datos, con el fin de evaluar los riesgos de fraude, como se indica en el Capítulo 1. En el desarrollo del modelo de riesgo también se han encontrado lagunas en los indicadores y bases de datos de riesgo de fraude, que si se abordan pueden ayudar a la IGAE a mejorar sus evaluaciones de riesgo de fraude, independientemente de la metodología específica que elija. En particular, la IGAE podría incorporar indicadores de comportamiento adicionales para evaluar riesgos de fraude en cada fase del ciclo de subvenciones, basándose en experiencias internacionales y literatura académica. Además, la IGAE podría recopilar datos empresariales y adoptar las metodologías descritas para realizar análisis de redes como un medio para identificar conflictos de interés, basándose en los ejemplos de compras públicas de este capítulo.

El capítulo también recoge una serie de consideraciones técnicas para la IGAE si decide adoptar un modelo de aprendizaje automático. Gran parte del trabajo pesado se ha realizado como parte de este proyecto piloto en términos de procesamiento y limpieza de datos. La IGAE tiene ahora un conjunto de datos de trabajo para usar en el análisis de riesgo de fraude, que ya es una mejora de lo que tenía disponible antes de este proyecto. Las características y limitaciones de los datos de la IGAE motivaron en gran parte la justificación para seleccionar el enfoque descrito. Si bien tiene limitaciones debido a la calidad de los conjuntos de datos de entrenamiento y las características de los datos de las subvenciones, la metodología se diseñó para minimizar los falsos positivos y los falsos negativos y, en general tiene un alto poder predictivo para identificar posibles fraudes en las subvenciones públicas de España. Si bien la implantación de este enfoque requiere capacidades adicionales, que se detallan en el Capítulo 1, la prueba de concepto muestra con éxito lo que es posible hacer con una inversión modesta, y proporciona una base para que la IGAE adopte una evaluación de riesgo de fraude verdaderamente basada en datos. El Capítulo 3 indaga más en cómo la IGAE puede mejorar la precisión del modelo interpretando datos adicionales que pueden usarse para detectar posibles fraudes.

Referencias

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45/1, pp. 5-32, https://link.springer.com/article/10.1023/a:1010933404324. [1]

Elkan, C. and K. Noto (2008), "Learning classifiers from only positive and unlabeled data", *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213-220, https://dl.acm.org/doi/10.1145/1401890.1401920. [2]

European Anti-Fraud Office (OLAF) (2011), *Compendium of Anonymised Cases*, https://ec.europa.eu/sfc/sites/default/files/sfc-files/OLAF-Intern-2011.pdf (accessed on 13 August 2021). [8]

European Commission (2020), *Report from the Commission to the European Parliament and the Council: 31st Annual Report on the protection of the European Union's financial interests — Fight against fraud - 2019*, https://ec.europa.eu/anti-fraud/sites/default/files/pif_report_2019_en.pdf (accessed on 13 August 2021). [9]

Fazekas, M., Ugale, G, & Zhao, A. (2019), *Analytics or Integrity: Data-Driven Decisions for Enhancing Corruption and Fraud Risk Assessments*, OECD Publishing, Paris, https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf. [7]

Fazekas, M. and I. Tóth (2016), "From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary", *Political Research Quarterly*, Vol. 69/2, pp. 320-334, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=H1FpS2AAAAAJ&citation_for_view=H1FpS2AAAAAJ:SP6oXDckpogC. [13]

Fazekas, M. and J. Wachs (2020), "Corruption and the network structure of public contracting markets across government change", *Politics and Governance*, Vol. 8/2, pp. 153-166, https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=PY3YH2kAAAAJ&citation_for_view=PY3YH2kAAAAJ:ZeXyd9-uunAC. [14]

Financial Fraud Enforcement Task Force (2012), *Reducing Grant Fraud Risk: A Framework For Grant Training*, https://www.oversight.gov/sites/default/files/oig-reports/Grant-Fraud-Training-Framework.pdf. [10]

James, G. et al. (2015), *Chapter 8*, https://link.springer.com/book/10.1007/978-1-4614-7138-7. [5]

Li, C. and X. Hua (2014), "Towards positive unlabeled learning", *International Conference on Advanced Data Mining and*, pp. 573–587. [3]

Lundberg, S. and S. Lee (2017), *A unified approach to interpreting model predictions*, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=ESRugcEAAAAJ&citation_for_view=ESRugcEAAAAJ:dfsIfKJdRG4C. [6]

Ministry of the Presidency of Spain (2021), *Resolution of May 26, 2020, of the Undersecretariat, which publishes the Agreement between the General Treasury of Social Security and the General Intervention of the State Administration, on the transfer of information*, https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-5748 (accessed on 4 July 2021). [15]

Mordelet, F. and J. Vert (2014), “A bagging SVM to learn from positive and unlabeled examples.”, *Pattern Recognition Letters*, Vol. 37, pp. 201-209, https://www.sciencedirect.com/science/article/pii/S0167865513002432. [4]

Unión Europea (2016), *Arachne, Be Distinctive*, http://www.ec.europa.eu/social/BlobServlet?docId=15317&langId=en (accessed on 13 August 2021). [12]

Wachs, J., M. Fazekas and J. Kertész (2020), “Corruption risk in contracting markets: a network science perspective”, *International Journal of Data Science and Analytics*, pp. 1-16, https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=PY3YH2kAAAAJ&citation_for_view=PY3YH2kAAAAJ:QIV2ME_5wuYC. [11]

Notas

[1] Para limpiar y fusionar los datos, se utilizó R 3.6.3 con el siguiente paquete: readxl, tidyverse (dplyr), flipTime, tibble, data.table. Para el modelado, se utilizaron R 3.6.3 y Python3 para las distintas fases de análisis. Para construir random forests en R, se utilizaron los paquetes randomForest y xgboost. Para el aprendizaje positivo sin etiquetar en las bibliotecas de Python3, se utilizaron pandas, numpy, baggingPU (módulo BaggingClassifierPU), sklearn.tree (módulos DecisionTreeClassifier, DecisionTreeRegressor, precision_score, recall_score, precision_score, train_test_split, RandomForestClassifier).

[2] Los nombres de estos conjuntos de datos incluyen BDNS_CONV_ACTIVIDADES, BDNS_CONV_ANUNCIOS, DNS_CONV_FONDOS_CON690, BDNS_CONV_OBJETIVOS_CON503, BDNS_CONV_TIPOBEN_CON590, BDNS_CONV_REGIONES_CON570, BDNS_CONV_INTRUMENTOS_CON560.

[3] Los nombres de estos conjuntos de datos incluyen BDNS_PROYECTOS, BDNS_PAGOS, BDNS_REINTEGRO, BDNS_DEVOLUCIONES.

[4] Los nombres de estos conjuntos de datos incluyen BDNS_INHABILITACIONES, BDNS_SANCIONES y BDNS_TERCERO_ACTIVIDADES_TER320

[5] El índice de precisión tan alto (95 %) se debe en gran parte al hecho de que la muestra está desequilibrada; es decir, la mayoría de los casos son negativos (no sancionados) y, por tanto, el modelo puede clasificar con relativa facilidad la mayor parte de la muestra como no sancionada. Sin embargo, es más difícil para el modelo predecir correctamente los casos sancionados, dado que son mucho menos frecuentes. Para este caso, la calificación de repetición es más útil para evaluar el rendimiento del modelo, ya que calcula el número de miembros de una clase que el clasificador identificó correctamente dividido por el número total de miembros de esa clase.