27.06.2022

Documento de conclusiones sobre el desarrollo y los resultados de las pruebas del proyecto "NDT - IA explicable en la gestión de Riesgos" presentado por Equifax Ibérica, S.L.

Departamento de Funciones Horizontales
Dirección General de Supervisión

1 Antecedentes

La Ley 7/2020, de 13 de noviembre, para la transformación digital del sistema financiero (en adelante, Ley 7/2020) regula un entorno controlado de pruebas que permite llevar a la práctica proyectos tecnológicos de innovación en el sistema financiero.

Con fecha 19 de febrero de 2021, Equifax Ibérica, S.L., en adelante el Promotor, presentó una solicitud para acceder al espacio controlado de pruebas conforme a un proyecto piloto, "NDT (IA explicable en la gestión de riesgos)," (en adelante, el "Proyecto"), cuyo objeto es presentar un modelo de calificación de riesgo crediticio, basado en algoritmos de Machine Learning[1] con restricciones de monotoneidad[2], denominado NeuroDecision Technology (en adelante "NDT"). Para el desarrollo del Proyecto se ha contado con un conjunto de datos de Banco Sabadell, participando éste como colaborador necesario.

El 14 de mayo de 2021 la Secretaría General del Tesoro y Financiación Internacional publicó en su sede electrónica la lista de proyectos que recibieron una evaluación previa favorable para acceder a dicho entorno de pruebas en la que figura incluido el Proyecto y se contempla que el Banco de España será la Autoridad Supervisora.

El 4 de agosto de 2021 el Banco de España y el Promotor suscribieron el protocolo de Pruebas (en adelante, el "Protocolo") en el que se recogen los términos en los que se realizarían las pruebas del Piloto presentado, al objeto de permitir al Promotor la realización –de manera controlada y delimitada– de las pruebas incluidas en el Proyecto.

Las pruebas previstas en el Protocolo se iniciaron con fecha 13 de septiembre de 2021 y finalizaron el 13 de abril de 2022.

[1] Machine Learning (aprendizaje automático) es un campo de la inteligencia artificial que aborda la construcción de modelos matemáticos para realizar predicciones y tomar decisiones de forma automatizada. Dichos modelos se construyen procesando datos y extrayendo patrones presentes en los mismos.

[2] Las restricciones de monotoneidad fuerzan al modelo a que las relaciones entre las variables explicativas y la variable objetivo sean siempre crecientes o siempre decrecientes.

Con fecha 11 de mayo de 2022, el Promotor remitió al Banco de España la Memoria, requerida por el apartado 1 del artículo 17 de la Ley 7/2020, con la evaluación de los resultados de las pruebas y del conjunto del proyecto piloto.

El apartado 3 del artículo 17 de la Ley 7/2020 establece que la autoridad que haya sido responsable del seguimiento de las pruebas elaborará un documento de conclusiones sobre su desarrollo y resultados. Dichas conclusiones se tendrán en cuenta a efectos de lo previsto en los artículos 25 (el Informe anual sobre transformación digital del sistema financiero elaborado por la Secretaría General del Tesoro y Financiación Internacional) y 26 (las autoridades supervisoras incluirán en su memoria anual un informe sobre la aplicación de la innovación de base tecnológica a sus funciones supervisoras). Las conclusiones se publicarán con las reservas necesarias de conformidad con lo previsto en la Ley 7/2020 y en los protocolos suscritos con los promotores.

En cumplimiento de lo establecido en el citado apartado 3 del artículo 17 de la Ley 7/2020, se elabora el presente Informe, en el que se recogen las conclusiones sobre el desarrollo de las pruebas y sus resultados.

2 Descripción del Proyecto

En el proyecto "NDT - IA explicable en la gestión de Riesgos" el Promotor ha presentado un modelo de calificación de riesgo crediticio, basado en algoritmos de Machine Learning con restricciones de monotoneidad, denominado NeuroDecision Technology (NDT), con el objetivo de demostrar que:

- Mejora la capacidad predictiva de los modelos tradicionales de regresión logística que utilizan las entidades.
- Mitiga el efecto caja negra, cumpliendo los requerimientos de transparencia y explicabilidad[3] que se exigen a la inteligencia artificial.

Para esto, se llevó a cabo una comparación cuantitativa de diferentes modelos de riesgo basados en algoritmos de Machine Learning y desarrollados sobre un conjunto de datos reales anonimizados proporcionado por el Banco Sabadell, participando éste como colaborador necesario.

El Promotor proporcionó la infraestructura tecnológica para el desarrollo del proyecto, consistente en una plataforma Big Data[4] de desarrollo analítico.

Para llevar a cabo esa comparación cuantitativa entre los diferentes modelos se llevó a cabo una separación de los datos en tres grupos de muestras: Entrenamiento[5], Validación[6] y Test[7]. A partir del análisis de creación de las muestras de entrenamiento y evaluación de

[3] Se refiere a la propiedad de un modelo de que se pueda entender fácilmente su lógica interna y saber por qué genera los resultados observados.
[4] Conjunto que contiene muchos tipos de datos provenientes de múltiples fuentes. El volumen de estos datos es muy elevado y crece a una velocidad muy alta. Por sus características, este conjunto de datos no puede tratarse con aplicaciones tradicionales.
[5] Conjunto de datos utilizado para que el modelo aprenda las características/patrones de los datos.
[6] Conjunto independiente al de entrenamiento, en el que se miden los resultados del entrenamiento y sirve para seleccionar el mejor conjunto de hiper parámetros para un modelo.
[7] Conjunto de datos utilizados para tener una métrica independiente y cuantificar el desempeño de un modelo después de terminar el proceso de entrenamiento. Este conjunto de datos se encuentra en la misma ventana temporal de la muestra de entrenamiento.

los modelos se generó una muestra adicional Out Of Time[8] que incluía los últimos doce meses del histórico del conjunto de datos (correspondiente al año 2019).

3 Desarrollo de las pruebas

3.1 Información remitida por el Promotor sobre el desarrollo de las pruebas

Las pruebas se llevaron a cabo entre los días 13 de septiembre de 2021 y 13 de abril de 2022, con una duración de 7 meses.

El proyecto se dividió en seis fases, cada una de ellas con una serie de tareas específicas, que se describen con detalle en la Memoria de evaluación de resultados proporcionada por el Promotor. Para llevar a cabo el marco de comparación entre distintas técnicas se entrenaron distintos modelos de Machile Learning: regresión logística, NDT, redes neuronales sin restricciones, XGboost con y sin restricciones de monotoneidad y Random Forest.

En la citada Memoria de evaluación de resultados, el Promotor señala que todas las tareas planificadas se habrían ejecutado sin incidencias. Por su parte, se habrían cumplido plenamente todos los criterios de éxito relacionados, a excepción de la prueba de comparación de resultados (correspondiente a la Fase V: comparativa entre modelos), en la que el criterio de éxito relacionado con la mejora en poder predictivo respecto de la regresión logística se habría cumplido parcialmente, tanto en NDT como en el resto de algoritmos de Machine Learning entrenados en la prueba.

Por tanto, la Memoria de evaluación de resultados identifica como objetivo no alcanzado plenamente la mejora de la capacidad predictiva, ya que las diferencias observadas en términos de poder predictivo entre la regresión logística y todos los algoritmos de Machine Learning, aun siendo en todos los casos superiores en estos últimos, fueron discretas, especialmente en el conjunto de test, e inferiores a lo observado en otros proyectos. El desempeño de todos los modelos, incluyendo la regresión logística, ha sido alto, reduciendo el posible margen de mejora de los algoritmos de Machine Learning.

La principal explicación que motiva estos discretos resultados, según el Promotor, es que las variables con más poder predictivo en este conjunto de datos mostraron una relación muy lineal respecto a la variable objetivo. Esto favoreció a la regresión logística frente a los otros algoritmos, que destacan cuando este comportamiento es no lineal.

Dado el resultado discreto obtenido, y con el objetivo de evidenciar el potencial de mejora de los modelos de Machine Learning respecto a la regresión logística, el Promotor ejecutó un ejercicio teórico adicional, entrenando los mismos algoritmos sin realizar el extenso preprocesado de ingeniería de variables. Tras analizar los resultados, el Promotor considera que se pone de manifiesto un claro aumento tanto de las diferencias entre modelos como en la precisión de los mismos.

De esta forma, como resultado de este proyecto, el Promotor considera que se ha podido establecer un marco metodológico para el entrenamiento y uso de modelos de Machine Learning en la gestión del riesgo de crédito y demostrar su potencial en términos de aumento de la inclusión financiera para los consumidores, y que gracias a su nivel de

[8] Conjunto de datos completamente fuera de la ventana temporal de los datos de entrenamiento. Es utilizado, entre otros aspectos, para cuantificar la generalización y estabilidad temporal de un modelo.

explicabilidad e interpretabilidad, alineado con el de la regresión logística, los modelos de Machine Learning con restricciones de monotoneidad, como es el caso de NDT, pueden ser considerados en el proceso de modelización de riesgo de crédito como métodos adicionales o alternativos a la regresión logística.

3.2 Seguimiento supervisor del desarrollo de las pruebas

A lo largo de las pruebas, el Banco de España y el Promotor mantuvieron una serie de reuniones de seguimiento periódicas para tratar, principalmente, su grado de avance.

A raíz de los resultados discretos de las pruebas descritos en el punto anterior, el Promotor realizó una prueba adicional, tipo laboratorio, en donde se limitó al máximo el tratamiento y la preselección de variables, proporcionando el mismo conjunto de datos en bruto a cada uno de los modelos basados en algoritmos de Machine Learning, aportando al Banco de España un informe adicional con el detalle de los criterios empleados y los resultados obtenidos.

3.3 Valoración supervisora del desarrollo de las pruebas

El proyecto se planificó en seis fases consecutivas durante las cuales se desarrollaron y analizaron los diferentes modelos. En este sentido, no ha habido ninguna incidencia de carácter crítico en una fase que haya impedido llevar a cabo las fases posteriores.

Sin embargo, la prueba en la que se cuantificaba la mejora predictiva y se calculaban las bandas de confianza ha tenido un resultado diferente del esperado y esto ha afectado a la consecución de, al menos, uno de los dos objetivos principales del proyecto.

A continuación, se aborda el grado de alcance de cada uno de los objetivos principales.

3.3.1 Mejora de las capacidades predictivas de los modelos logísticos

El primer objetivo del proyecto consistía en demostrar que el modelo NDT era capaz de conseguir mejores capacidades predictivas que una regresión logística a la hora de predecir el incumplimiento[9]. Este objetivo es interesante dado que la regresión logística es desde hace décadas una de las técnicas más ampliamente utilizadas para este propósito.

Las pruebas relacionadas con la consecución de este hito son:

- En la fase IV, se evaluó la capacidad predictiva de los modelos y se observó una mejora muy limitada del poder predictivo de NDT con respecto a la regresión logística en las dos muestras de test consideradas (la mejora fue menor al 1% en términos del AUC-ROC[10]).
- En la fase V se calculó la banda de confianza de las métricas de poder predictivo y se observó que presentaban un importante solapamiento en las dos muestras de test consideradas.

[9] Impago superior a 90 días

[10] La curva ROC (Receiver Operating Characteristic) es una figura que muestra el desempeño de un modelo de clasificación en todos los umbrales de clasificación; está relacionada con la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). El AUC-ROC (o área bajo la curva ROC) toma valores entre 0 y 1. Un modelo con el 100% de predicciones incorrectas tendrá un AUC-ROC igual a 0, mientras que en el caso de un modelo perfecto su valor será igual a 1.

Por tanto, la limitada ganancia de poder predictivo de NDT observada no es estadísticamente significativa y este objetivo no ha sido alcanzado con éxito.

3.3.2 Mejora de la explicabilidad de los modelos de aprendizaje automático

El segundo objetivo del proyecto consistía en demostrar que el modelo NDT es más interpretable que otros modelos de aprendizaje automático en el contexto del modelado del riesgo de crédito. Este objetivo es interesante dado que los modelos de aprendizaje automático, a los que se les presupone mejores prestaciones en términos de capacidad predictiva que a las regresiones logísticas, sufren de una gran opacidad, ya que resulta difícil descifrar la lógica interna de los mismos.

Las pruebas relacionadas con la consecución de este hito se desarrollaron durante la fase V. Estas pruebas evidencian cómo los modelos con restricciones de monotoneidad presentan mejores características en términos de explicabilidad que los modelos de aprendizaje automático sin este tipo de restricciones. Sin embargo, es importante notar que NDT no presenta ventajas en términos de explicabilidad frente al XGBoost con restricciones de monotoneidad.

Por tanto, este objetivo se puede considerar alcanzado, aunque con ciertas limitaciones.

Por otra parte, no se han materializado riesgos durante las pruebas.

4 Próximos pasos

4.1 Información remitida por el Promotor

El Promotor indica en la Memoria de evaluación de resultados que en este Proyecto no aplica incorporar una descripción de los planes de resolución del piloto y cómo se decomisará el servicio, puesto que es una prueba de laboratorio donde desarrollaron los modelos en un espacio controlado, pero que no han sido implantados en un proceso de negocio.

En cuanto al tratamiento de datos personales, todos los datos presentados fueron anonimizados al comienzo de las pruebas y previo al envío de los mismos a los entornos de Equifax, por parte de Banco Sabadell, por lo que no se utilizó ninguna variable con datos personales.

El Promotor señala que, gracias a los resultados de la prueba, pretende:

- Incorporar NDT como algoritmo adicional en sus desarrollos internos de modelos de calificación genéricos.
- Potenciar la oferta de proyectos de desarrollo de modelos de calificación para clientes, incorporando tanto NDT como otros algoritmos de Machine Learning, dando un especial énfasis en el marco de explicabilidad e interpretabilidad que brinda un modelo como NDT.

Asimismo, con el objetivo de fomentar la innovación y el uso de nuevas tecnologías en el mundo financiero, el Promotor indica que presentará las conclusiones principales del uso de este tipo de algoritmos en distintos foros y a distintos interlocutores, tanto instituciones públicas (reguladores, supervisores, institutos de investigación) como privadas (bancos, aseguradoras, empresas tecnológicas y Fintechs).

4.2 Valoración supervisora sobre los siguientes pasos del Proyecto

En relación a las pretensiones indicadas por el Promotor, no se identifican factores de riesgo que pudieran suponer un condicionante por parte del Banco de España, al tiempo que no se precisaría de modificaciones normativas para la puesta en producción del proyecto o para favorecer la innovación financiera.

Para llevar a cabo el Proyecto fuera del entorno del *Sandbox*, el Promotor no hará uso de la pasarela de acceso a la actividad a la que se refiere el artículo 18 de la Ley 7/2020, ya que no es necesario solicitar autorización para implantar este Proyecto.

Se advierte de que el Banco de España no ha llevado a cabo una valoración del cumplimiento del principio de responsabilidad proactiva en el tratamiento de datos personales, toda vez que dicha valoración excede del ámbito competencial de esta Institución.

5 Barreras regulatorias identificadas por el Promotor

El Promotor señala que no existe ninguna barrera regulatoria, más allá de que, desde un punto de vista de uso de la Inteligencia Artificial para la toma de decisiones automatizadas, el Reglamento General de Protección de Datos (RGPD) en sus Considerandos 63 y 71, así como en su artículo 22, limita y establece derechos con relación a que los sujetos de los datos no sean sometidos a decisiones exclusivamente automatizadas que tengan efectos jurídicos o que afecten significativamente al interesado. Igualmente, esta normativa otorga a los sujetos el derecho a que puedan conocer la lógica implícita, así como a oponerse en cualquier momento a la elaboración de perfiles.

La elaboración de perfiles de forma automática se incluye en este marco de decisiones automatizadas. Esta norma podría constituir una barrera regulatoria para aplicar cualquier modelo que no sea transparente y suficientemente supervisado por un humano en el momento de ser utilizado para la toma de decisiones en la gestión del riesgo de crédito. En este sentido, consideran que soluciones como NDT aportan un marco de transparencia, dado su nivel de explicabilidad e interpretabilidad, que facilitan el entendimiento de la decisión del modelo de un consumidor.

6 Conclusiones

6.1 Conclusiones remitidas por el Promotor

En su Memoria de evaluación de resultados, el Promotor señala que, durante el desarrollo de las pruebas, los algoritmos Machine Learning demostraron tener el potencial de superar el desempeño de la regresión logística, traduciéndose en un beneficio tanto para las entidades financieras como para los consumidores, ya que modelos desarrollados utilizando algoritmos de Machine Learning con restricciones de monotoneidad, como NDT, aumentan la inclusión financiera frente a la regresión logística[11], porque habría un mayor

[11] El Promotor llevó a cabo un experimento simulado a partir de datos históricos para evaluar el número de préstamos que se otorgarían en función del modelo utilizado para evaluar el riesgo de crédito de los prestatarios. En este experimento se concluía que al utilizar NDT el número de préstamos otorgado sería mayor que si se hubiese utilizado una regresión logística.

número de personas que tendrían acceso a la financiación, garantizando asimismo un marco de explicabilidad e interpretabilidad alineado.

Como lecciones aprendidas, el Promotor indica que los conjuntos de datos que incluyen información comportamental[12] aumentan el potencial de los modelos basados en regresión logística, reduciendo los posibles márgenes de mejora de otros modelos. Además, sería conveniente explorar nuevas variables que puedan ser utilizadas por este tipo de algoritmos, o aplicarlos a tipos de problemas en los que la regresión logística tiene un menor desempeño; por ejemplo, en problemas de admisión. Finalmente, indica que los resultados obtenidos no serían extrapolables a todas las carteras.

Por otra parte, considera que el marco de explicabilidad en este tipo de modelos puede ser comparable al de modelos basados en regresiones logísticas, aportando una transparencia que se traduce en protección para el consumidor y cliente de servicios financieros, al tiempo que facilita su uso para los gestores de riesgo en su fijación de estrategias y uso diario y fomenta la innovación desde un punto de vista del uso de la Inteligencia Artificial en la gestión de riesgos.

También favorecería la igualdad de género, al disminuir posibles sesgos de tipo ético como consecuencia de una mayor automatización del proceso de concesión.

Por último, respecto a la evolución internacional, considera que el conocimiento desarrollado durante el proyecto podría contribuir a rellenar el gap de tecnología e innovación que la Unión Europea tienen con otros países como EEUU y Canadá.

6.2 Conclusiones supervisoras

El primer objetivo, probar que NDT presenta mejores capacidades predictivas que los modelos logísticos, no ha sido alcanzado con éxito. Se ha observado únicamente una mejora marginal que no es estadísticamente significativa.

El segundo objetivo, probar que mitiga el efecto de caja negra de los modelos de aprendizaje automático, ha sido alcanzado en gran medida, aunque con reservas. Se ha visto que el modelo NDT es más interpretable que las técnicas de aprendizaje automático sin restricciones analizadas (red neuronal, XGBoost, Random Forest), pero resulta equivalente al modelo XGBoost con restricciones de monotoneidad.

En todo caso, a efectos de un posible uso efectivo del modelo propuesto en el proceso de concesión crediticia debe tenerse en cuenta el principio básico de transparencia que sustenta el RGPD, en el sentido de que no se lleven a cabo decisiones exclusivamente automatizadas que tengan efectos jurídicos o que afecten significativamente a los interesados, así como que se garantice que las decisiones que se deriven del modelo sean explicadas a las personas de forma clara y sencilla.

Con respecto al impacto del uso de NDT en la inclusión financiera, la prueba realizada no permite extraer conclusiones suficientemente sólidas, ya que:

- Se basa en un único escenario hipotético cuyas asunciones no están suficientemente justificadas.

[12] El Promotor define variables comportamentales como aquellas que recogen el comportamiento de un consumidor con una entidad financiera una vez que éste ya es cliente

- El incremento de las concesiones observado en este escenario va asociado a un aumento importante de las posiciones morosas.

Conforme a lo dispuesto en el artículo 5 de la Ley 7/2020, el Proyecto debía aportar potencial utilidad o valor añadido. A este respecto, el Supervisor concluye que el uso de algoritmos de Machine Learning con restricciones de monotoneidad puede representar una ventaja para las entidades financieras, al poder disponer de herramientas alternativas a los modelos logísticos que cuentan con un marco de explicabilidad e interpretabilidad alineado a los mismos.

Sin embargo, la supervisión de estos modelos desarrollados utilizando algoritmos de Machine Learning es mucho más compleja que la que se lleva a cabo sobre los modelos logísticos. Por lo tanto, puede concluirse que la innovación probada en este Proyecto no proporciona mecanismos para el mejor ejercicio de la función supervisora, por lo que no sería necesaria la inclusión de la evaluación del Proyecto en el informe para la Memoria de Supervisión al que hace referencia el artículo 26 de la Ley 7/2020. Si bien es cierto que estos no eran objetivos buscados en el planteamiento del Proyecto por parte del Promotor.

Por delegación de la Comisión Ejecutiva
B.O.E. de 27.12.2019

[F] Mª MERCEDES OLANO LIBRAN

Firmado digitalmente por [F] Mª MERCEDES OLANO LIBRAN

Mercedes Olano
Directora General de Supervisión