Newsletter trimestral

CÁTEDRA

iDANAE

INTELIGENCIA · DATOS · ANÁLISIS · ESTRATEGIA

3T19

Interpretabilidad de los modelos de Inteligencia Artificial

POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

MS ManagementSolutions
Making things happen

Análisis de metatendencias

La Cátedra iDanae (inteligencia, datos, análisis y estrategia) en Big Data y Analytics, que surge en el marco de colaboración de la Universidad Politécnica de Madrid (UPM) y Management Solutions, tiene el objetivo de promover la generación de conocimiento, difusión y transferencia de tecnología, y fomento de la I+D+i en el área de Análisis de Datos.

En este contexto, una de las líneas de trabajo que desarrolla esta Cátedra es el análisis de las metatendencias en el ámbito de Analytics. Una metatendencia se puede definir como un concepto o ámbito de interés de un campo de conocimiento determinado que requerirá de inversión y desarrollo por parte de gobiernos, empresas y la sociedad en general en un futuro próximo, por ser generador de valor.

En la detección de las metatendencias es importante analizar los proyectos de inversión públicos y privados, así como los elementos destacados por organizaciones, empresas y otros grupos de interés relacionados. Desde esta Cátedra se mantendrá una vigilancia activa a través del observatorio y seguimiento de distintas fuentes, como el resultado de los grupos de trabajo europeos en Analytics[1], los planes estratégicos del Gobierno de los Estados Unidos sobre la estrategia de investigación y desarrollo en inteligencia artificial[2], y otros análisis y publicaciones relevantes a nivel internacional.

Con un fin educativo y divulgativo, fruto de esta vigilancia se mantiene una lista actualizada de temáticas que se irán desarrollando a través de la publicación de un informe trimestral, con el objetivo de aportar una visión de una determinada tendencia o tema de interés. Entre las temáticas que se irán abordando se destaca a continuación un ejemplo de algunas consideradas relevantes:

- Interpretabilidad de modelos.
- Implicaciones éticas, legales y sociales de la Inteligencia Artificial.
- Predictibilidad y modelizabilidad: límites de lo que se puede modelizar y de la aportación de un modelo.
- Datos y técnicas: un enfoque estratégico de la modelización y estrategia de los datos.
- IA práctica: un enfoque de desarrollo.
- Data augmentation y data democratization
- Tratamiento de datos: Inteligencia Artificial y leyes de protección de datos.

En este primer informe trimestral se ha desarrollado el concepto de interpretabilidad de modelos.

[1]Big Data Value Association, EU Robotics.
[2]NSTC Committee on AI 2019.

Concepto de interpretabilidad

De acuerdo con la BDVA y EU Robotics[3], el término Inteligencia Artificial (IA) se usa como un término global que cubre la inteligencia tanto digital como física, datos y robótica, y tecnologías inteligentes relacionadas. El desarrollo de la inteligencia artificial incluye modelos y técnicas de distinto nivel de complejidad, lo que puede reducir la capacidad de que un ser humano comprenda los modelos subyacentes o los resultados obtenidos.

Los conceptos de explicabilidad[4] e interpretabilidad aparecen como respuesta a la búsqueda de la comprensión de los modelos de inteligencia artificial. Ambos términos se encuentran estrechamente relacionados, si bien aparecen en la literatura con diferente significado.

Por ejemplo, en (Gandhi, 2019) se establece una clara distinción entre interpretabilidad y explicabilidad:

- La interpretabilidad se refiere al grado en que se puede observar una causa y su efecto dentro de un sistema, es decir, en qué medida se puede explicar el resultado de un modelo dado un cambio en los datos de entrada.

- La explicabilidad se refiere al grado en el que el comportamiento del modelo se puede explicar en términos humanos, considerando tanto el resultado como todo el proceso de la toma de decisión. Esto incluye:

    - Explicar los datos utilizados y cómo se obtienen los resultados.
    - Explicar cómo se producen los outputs a través de los inputs.
    - Explicar la intención con la que el sistema afecta a las partes involucradas.

No obstante, otros autores aportan diferentes definiciones sobre interpretabilidad y explicabilidad. Como ejemplos:

- Un sistema de inteligencia artificial se puede definir como interpretable si sus operaciones pueden ser entendidas por el ser humano, ya sea por introspección o a través de algún tipo de explicación (Or Biran, 2017)[5].

- Interpretar significa explicar en términos entendibles. En el contexto de sistemas de Machine Learning, se define la interpretabilidad como la habilidad de explicar en términos comprensibles para un ser humano (Finale Doshi-Velez, 2017).

- Se define como interpretable a los modelos de Machine Learning que extraen información relevante de las relaciones contenidas en los datos, entendiendo información relevante si proporciona comprensión a una particular audiencia y guían la comunicación, los actos o los descubrimientos (Murdoch, 2019).

[3]BDVA, EU Robotics.
[4]Aunque el término explicabilidad no aparece en el diccionario de la RAE, explicabilidad es el sustantivo adecuado para aludir a la cualidad o condición de lo explicable, es decir, de lo que se puede explicar, según la propia Academia.
[5]Nótese que esta definición estaría más alineada con el concepto de explicabilidad antes mencionado.

Por tanto, el concepto de explicabilidad podría entenderse como un concepto más amplio, con un objetivo más ambicioso que la interpretabilidad. En este documento se persigue desarrollar el concepto de interpretabilidad entendida como todo modelo de inteligencia artificial donde sea posible comprender las razones por las que se ha realizado una predicción concreta y las interacciones establecidas entre las diferentes variables.

La búsqueda de la interpretabilidad en modelos de Machine Learning surge a partir del creciente uso de modelos complejos (Hastie, Tibshirani, & Friedman, The elements of statistical learning, 2009), como por ejemplo las redes neuronales, donde la complejidad de las relaciones y cálculos, la estructura y el volumen de parámetros que se han de estimar dificultan su interpretación. Así, el objetivo de los análisis para lograr la interpretabilidad se centran en el entendimiento tanto de dichas relaciones como de los resultados y conclusiones.

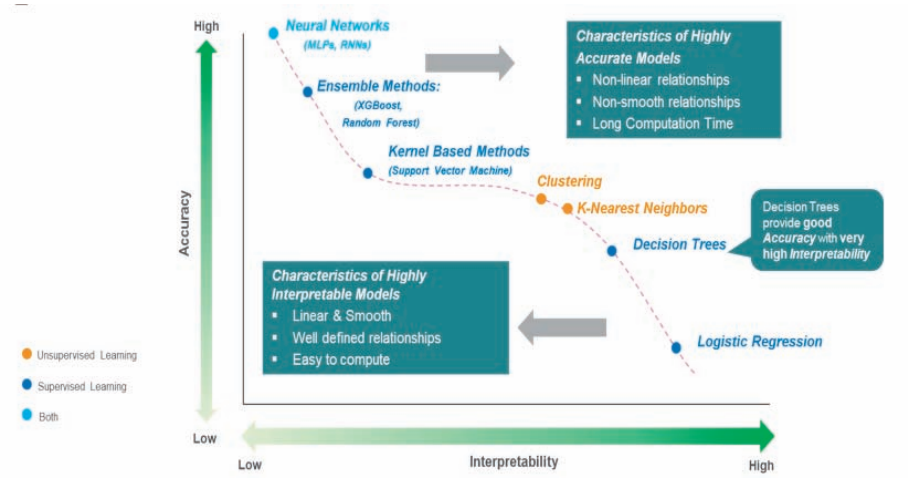
Sin embargo, en términos generales existe una relación inversa entre interpretabilidad y poder predictivo del modelo: un modelo más simple supone una mayor capacidad de interpretación, pero suele implicar una capacidad predictiva menor y viceversa (ilustración 1).

En general, en el entorno del desarrollo de un proyecto de Inteligencia Artificial se deberá analizar este trade-off, considerando elementos como el objetivo del proyecto, el uso que se hará del resultado y las métricas de éxito, entre otros. Por ejemplo, cuando el carácter interpretable es fundamental en un proyecto, el enfoque tradicional para lograrlo se ha basado en limitar la tipología de técnicas de aprendizaje automático, circunscribiendo las alternativas a aquellos algoritmos más sencillos, interpretables desde el punto de vista de su estructura o de su entrenamiento, o bien en la simplificación de modelos complejos, de forma que la comprensión de su funcionamiento sea más accesible. Este enfoque suele resultar ventajoso cuando las relaciones en los datos son simples (lineales). Las principales técnicas utilizadas son los modelos lineales generalizados, árboles de decisión, clasificadores bayesianos ingenuos o k-vecinos más próximos, consiguiendo una elevada capacidad discriminante y/o poder predictivo.

No obstante, si no se quiere renunciar al uso de técnicas menos interpretables, se puede buscar la interpretabilidad a partir del análisis de los outputs del modelo. El objetivo en este caso consiste en extraer información a través de diversos métodos, logrando conseguir un entendimiento del modelo de forma indirecta. Por ejemplo, se pueden analizar pequeños cambios en el modelo o los datos de entrenamiento y observar el resultado, o construir modelos interpretables más sencillos que expliquen el modelo complejo (aunque dichos modelos puedan tener un rendimiento peor). En este documento se exponen algunas técnicas de interpretabilidad basadas en estos tipos de análisis[6].

[6]Estas técnicas están relacionadas con las denominadas "técnicas post-hoc".

Ilustración 1. Interpretabilidad vs Capacidad Predictiva



Fuente: Sharayu Rane (2018).

Interpretabilidad basada en el análisis

Las técnicas de interpretabilidad basada en el análisis toman como datos de entrada el modelo entrenado (junto con las predicciones ya realizadas), y extraen información sobre las relaciones que el modelo ha aprendido, ya sea entrenando modelos interpretables con las predicciones del modelo no interpretable (conocido como la construcción de un modelo subrogado), o realizando perturbaciones en los datos de entrada y estudiando la reacción del modelo. Son métodos especialmente útiles cuando se trabaja con relaciones complejas que necesitan modelos black-box para lograr una exactitud predictiva razonable. Además, estos métodos son aplicables a diferentes modelos, pudiendo servir como medida para comparar entre las predicciones de diferentes modelos.

A continuación, se detallan tres de estos métodos por ser ampliamente utilizados: el gráfico de dependencias parciales (PDP, por sus siglas en inglés), LIME (Local Interpretable Model-agnostic Explanation) y SHAP (SHapley Additive exPlanations). Otros métodos no tratados en este informe incluyen SKATER (Choudhary, Kramer, & team, 2018), ELI5 (Mikhail Korobov, s.f.) o DALEX (Biecek, 2018), entre otros.

Para ilustrar estos métodos se ha realizado un caso práctico utilizando una base de datos de 45.211 entradas relacionadas con campañas de marketing de un banco portugués para vender un depósito bancario[7]. La campaña fue realizada telefónicamente, pudiendo haberse realizado varias llamadas al mismo cliente para contratar o no el producto. El objetivo es predecir si el cliente contratará el depósito (variable "y"). Existen variables relacionadas con los datos personales y bancarios del cliente (edad, trabajo, estado civil, educación, si ha impagado algún crédito, si tiene préstamos personales o hipotecarios y su balance anual promedio), con la campaña actual (medio por el que se han comunicado, día y mes del último contacto y duración del último contacto, número de veces que se ha contactado al cliente en esta campaña) y con campañas anteriores (el número de días desde que se le contactó por última vez en la campaña anterior, el número de contactos que se realizaron y el resultado de la campaña). En las ilustraciones 2-5 se realiza una pequeña exploración visual de algunas de las variables mencionadas.

Distribución de frecuencias de la edad.

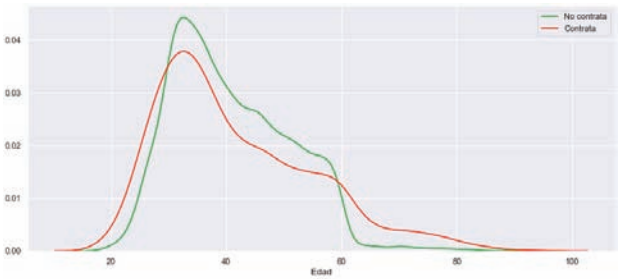


Ilustración 2. Distribución de frecuencias de la edad para los clientes que han contratado el depósito y los que no.

Target plot for feature "age".
Average target value through different feature values.

Ilustración 3. Contratos del depósito y probabilidad en función de la edad.

Target plot for feature "balance"
Average target value through different feature values.

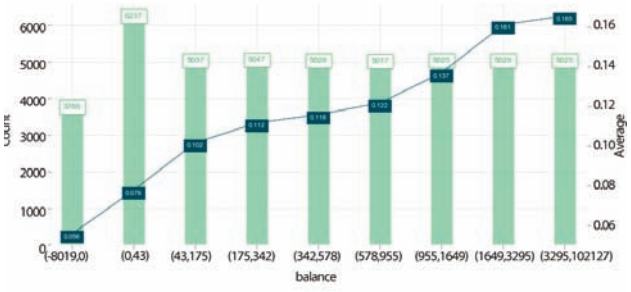


Ilustración 4. Contratos del depósito y probabilidad en función del balance.

[7]Los detalles de la base de datos pueden encontrarse en (Moro y otros, 2011).

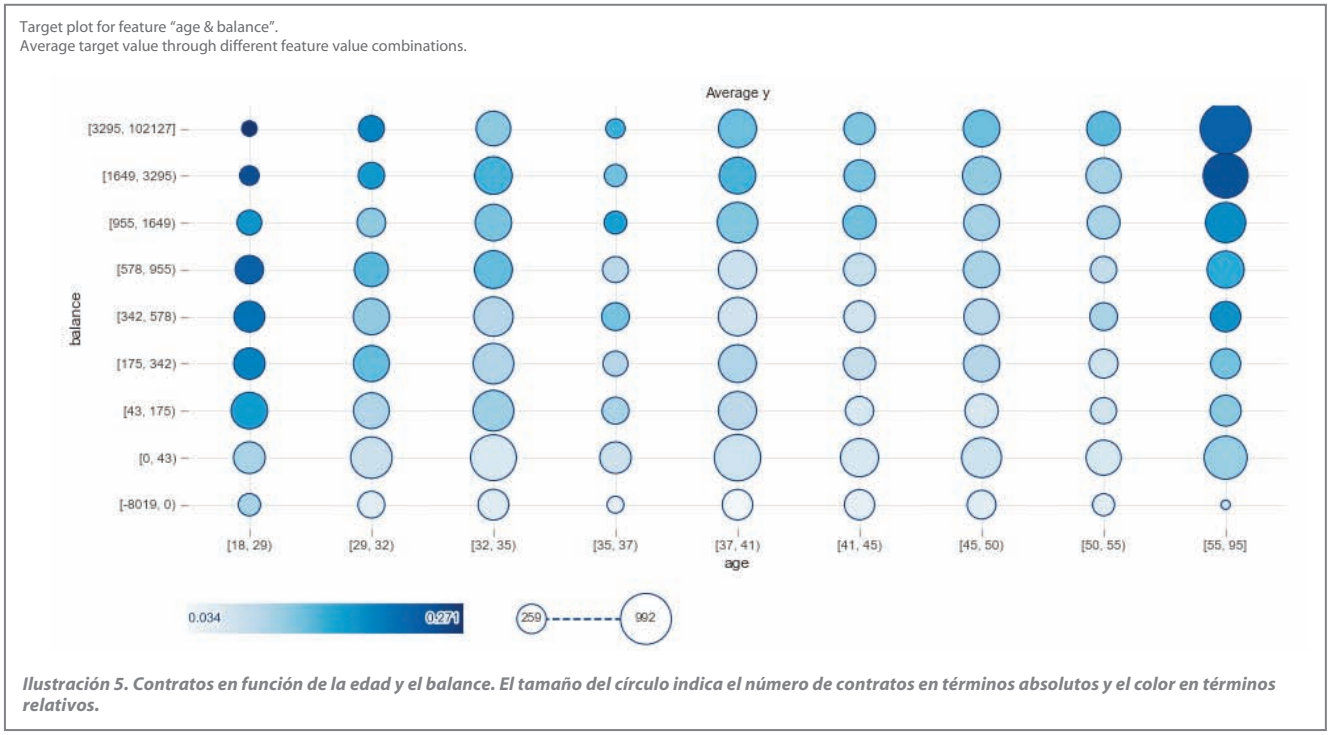


Ilustración 5. Contratos en función de la edad y el balance. El tamaño del círculo indica el número de contratos en términos absolutos y el color en términos relativos.

Gráfico de Dependencias Parciales (PDP)

Los gráficos PDP (Friedman, 2001) muestran el efecto marginal de una o dos variables independientes en la predicción realizada, así como el tipo de relación entre variables independientes y dependientes.

Este modelo separa el conjunto de variables independientes en dos subconjuntos: un subconjunto S con un número determinado de variables, y otro C con el resto de variables. Este método funciona marginalizando la predicción del modelo utilizado sobre la distribución de las variables en C, de forma que el gráfico muestre la dependencia entre las variables en S y la predicción. Si se elige S de forma que contenga solamente una o dos variables, entonces es posible representar gráficamente el resultado, obteniendo un gráfico de dependencias parciales.

Los gráficos de esperanza condicional individual, o gráficos ICE (Goldstein, 2015), funcionan de forma similar a los gráficos PDP, pero tratando de forma individual cada observación, en lugar de trabajar con los promedios. Esto significa que, si se tienen N observaciones, en dicho gráfico se van a tener N líneas dibujadas, en lugar de una sola línea promediada como se señalaba en PDP. Este método funciona mejor que PDP cuando existen interacciones (situación en la que dos o más variables actúan entre sí para generar un nuevo efecto) entre las variables independientes.

Caso práctico: En la ilustración 6 se puede ver un gráfico ICE y PDP para la variable balance y en la ilustración 7 para la variable Duration (duración de la última llamada). Las líneas azules son los gráficos ICE para una submuestra de 300 predicciones, mientras que la línea amarilla muestra el gráfico PDP, siendo el promedio de todos los gráficos ICE. Como se puede observar, mayores balances suponen un aumento en la probabilidad en que el cliente contrate el depósito, mientras que una mayor duración de la llamada también implica una mayor probabilidad de contratación.

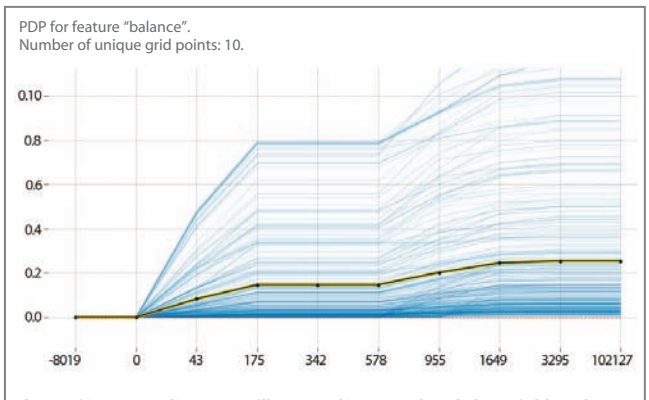


Ilustración 6. PDP (línea amarilla) e ICE (líneas azules) de la variable Balance.

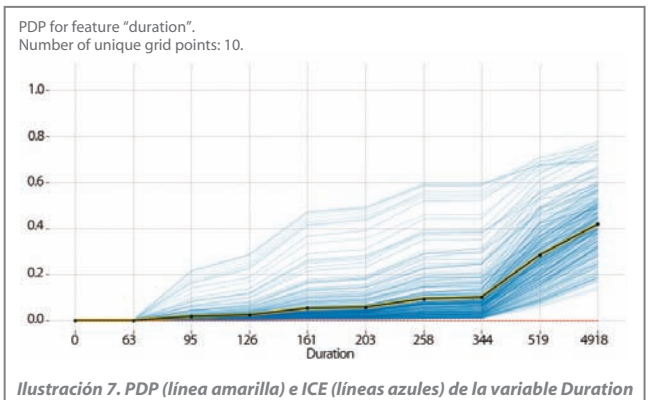


Ilustración 7. PDP (línea amarilla) e ICE (líneas azules) de la variable Duration (duración de la llamada).

LIME

LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, & Guestrin, 2016) es un método local que comprueba qué ocurre con las predicciones de un modelo cuando los datos introducidos varían. Este no es un modelo puramente transparente, ya que proporciona la explicación después de que se haya tomado una decisión. LIME genera unos nuevos datos compuestos por datos modificados y las predicciones generadas por el modelo. Sobre estos nuevos datos se entrena un modelo interpretable (como los mencionados anteriormente: modelos lineales, árboles de decisión, etc.). Este modelo interpretable debe ser una buena aproximación de las predicciones de manera local.

Se define la explicación ξ como un modelo $g \in G$, donde G es una clase de modelos potencialmente interpretables. Como no todos los modelos son lo suficientemente simples como para poderse interpretar con facilidad en todas las situaciones, se define una medida de complejidad $\Omega(g)$ (para árboles de decisión, por ejemplo, es la profundidad del árbol; para regresiones lineales, el número de pesos no nulos). Se debe establecer también una medida de proximidad $\pi_x(z)$ que establezca una distancia entre la observación x y la observación z para poder definir la localidad de x. Esto permite definir $L(f,g,\pi_x)$ como una medida de la fidelidad en la aproximación de f por g en la proximidad de x definida por π_x. Por tanto, para conseguir una interpretación de la predicción de x y teniendo una buena aproximación local del modelo se debe minimizar $L(f,g,\pi_x)$ a la vez que se debe mantener $\Omega(g)$ lo suficientemente bajo para que sea interpretable por humanos.

Caso práctico: En la ilustración 8 se puede ver la información facilitada por LIME para una de las predicciones. En este caso concreto, la predicción es que dicho cliente no va a contratar el depósito, y las principales razones por las que el modelo predice eso es debido, principalmente, a los valores que tienen las variables duration (duración de la última llamada), pdays (número de días desde que fue contactado en la campaña anterior, siendo -1 el valor asignado en caso de no haber sido contactado), previous (contactos realizados en la campaña anterior) y contact (forma por la que se contacta: desconocido, teléfono o móvil).

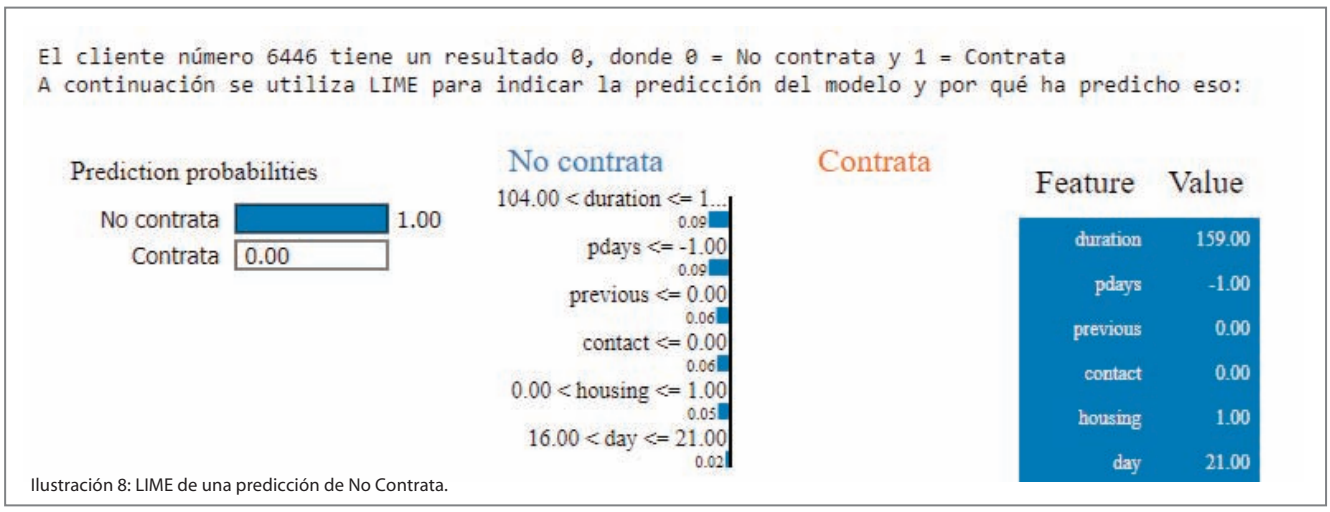


Ilustración 8: LIME de una predicción de No Contrata.

SHAP

SHAP (Scott Lundberg, 2017) es un método local que utiliza la teoría de juegos cooperativos para interpretar las predicciones realizadas por un modelo. En este método se calculan los valores de Shapley (Shapley, 1953), en los que las variables independientes se interpretan como jugadores que colaboran para recibir el payout, que en este caso es la predicción concreta realizada por el modelo menos el valor promedio de todas las predicciones. Los jugadores se "reparten" el payout en función de su contribución, y este reparto viene calculado por los valores de Shapley.

Este método requiere reentrenar el modelo en todos los posibles subconjuntos $S \subseteq F$, siendo F el conjunto de todas las variables independientes, asignando a cada variable un valor de importancia que representa el efecto de dicha variable en la predicción del modelo. Para ello, se entrena un modelo $f_{(S \cup \{i\})}$ con la variable presente y otro modelo f_S sin la variable presente. La predicción de ambos modelos se compara para la entrada concreta x_S que se quiere predecir. Como el efecto de negar una variable depende de otras variables en el modelo, estas diferencias se calculan para todos los posibles subconjuntos posibles $S \subseteq F \setminus \{i\}$, resultando los valores de Shapley en la media ponderada de todas las posibles diferencias.

Caso práctico: En la ilustración 9 se puede observar la misma predicción explicada anteriormente por LIME. Como se puede observar, las variables contact, housing y pdays son las mayores contribuyentes para aumentar la probabilidad. Cabe destacar que SHAP no muestra la probabilidad de default sino la contribución promedio de una variable a la predicción (cuánto se modifica la predicción de forma promedio considerando todos los subconjuntos sin contener la variable y conteniendo la variable). No es una probabilidad ni un valor que indica la diferencia en la predicción si se elimina esa variable.

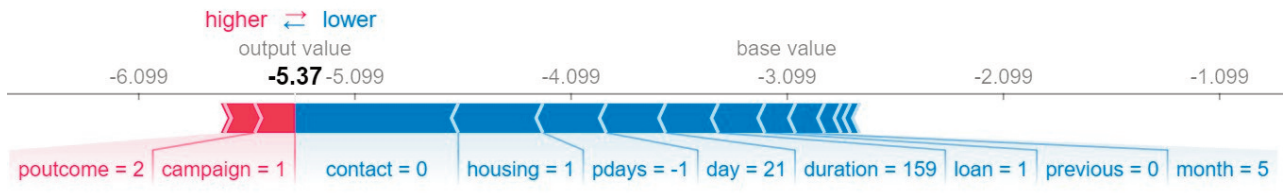


Ilustración 9. Predicción explicada por SHAP.

Comparación entre PDP, LIME y SHAP

En la siguiente tabla, se muestra una comparación entre los tres métodos señalados.

	PDP	LIME	SHAP
CONCEPTO	Marginaliza todas las variables menos una para ver cómo afecta en la predicción variaciones en la variable.	Aproxima pequeñas variaciones de la predicción, de forma local, a modelos interpretables.	Supone la predicción como un juego cooperativo entre las variables en las que se reparten un payout (la predicción realizada) de forma ponderada a su contribución.
VENTAJAS	La interpretación es muy intuitiva y causal, además de ser sencillo de implementar.	No es necesario utilizar todas las variables y funciona para textos e imágenes. Tiene capacidad predictiva para entornos de la variable.	Los fundamentos matemáticos detrás de SHAP lo convierten en una teoría de interpretabildiad sólida.
DESVENTAJAS	Sólo permite explicar una o dos variables a la vez. Se asume independencia entre las variables.	Es complicado definir el entorno de las variables. Las explicaciones entre pequeñas variaciones de las variables pueden ser considerablemente distintas.	Supone un gran coste computacional y el resultado puede malinterpretarse. Siempre se utilizan todas las variables y no tiene capacidad predictiva.

Tabla 1: Comparación entre PDP, LIME y SHAP.

Referencias

BDVA, EU Robotics: "Strategic Research, Innovation and Deployment Agenda for an AI PPP". Consultation Release. Junio 2019.

Biecek, P. (2018). DALEX: explainers for complex predictive models. The Journal of Machine Learning Research.

Choudhary, P., Kramer, A., & team, d. (2018). Skater: Model Interpretation Library. https://zenodo.org/record/1198885.

Finale Doshi-Velez, K. B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. Retrieved from Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics , 1189-1232.

Gandhi, P. (2019). Explainable Artificial Intelligence. Retrieved from KD Nuggets: https://www.kdnuggets.com/2019/01/explainable-ai.html

Goldstein, A. e. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24.1, 44-65.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer.

M. Lichman, U. o. (s.f.). UCI Machine Learning Repository . Obtenido de UCI Machine Learning Repository : http://archive.ics.uci.edu/ml

Mikhail Korobov, K. L. (s.f.). ELI5. Obtenido de Github: https://github.com/TeamHG-Memex/eli5

Murdoch, W. J. (2019). Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

Moro y otros, 2011: S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, Octubre, 2011. EUROSIS.

NSTC Committee on AI 2019: Committee on AI of the National Science and Technology Council: The national artificial intelligence research and development strategic plan: 2019 update. Executive Office of the President of the United States. June 2019.

Or Biran, C. C. (2017). Explanation and Justification in Machine Learning: A Survey. IJCAI 2017 Workshop on Explainable Artificial Intelligence.

Ribeiro, M. T., Singh, S., & Guestrin, a. C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM.

Scott Lundberg, S.-I. L. (2017). A Unified Approach to Interpreting Model Predictions. NIPS.

Shapley, L. S. (1953). A value for n-person games. Contributions to the Theory of Games.

Sharayu Rane (2018): "The Balance: Accuracy vs Interpretability". Towards Data Science. 2018.

Autores

Ernestina Menasalvas (UPM)
Alejandro Rodríguez (UPM)
Manuel Ángel Guzmán (Management Solutions)
Segismundo Jiménez (Management Solutions)
Carlos Alonso (Management Solutions)

UNIVERSIDAD POLITÉCNICA DE MADRID

La Universidad Politécnica de Madrid es una Entidad de Derecho Público de carácter multisectorial y pluridisciplinar, que desarrolla actividades de docencia, investigación y desarrollo científico y tecnológico.

www.upm.es

Management Solutions es una firma internacional de consultoría, centrada en el asesoramiento de negocio, finanzas, riesgos, organización, tecnología y procesos, que opera en más de 40 países y con un equipo de más de 2.500 profesionales que trabajan para más de 900 clientes en el mundo.

www.managementsolutions.com

Para más información visita

blogs.upm.es/catedra-idanae/