



SKILLING
CENTER

TECMILENIO



Inteligencia artificial y machine learning

Interpretación de modelos
de aprendizaje automático





Las empresas buscan tomar decisiones efectivas y eficientes en el presente para que impacte de forma positiva en el futuro. Los indicadores financieros que apoyan una decisión son considerados con base en predicciones. De ahí la necesidad de considerar modelos predictivos que puedan relacionar en el presente, situaciones que pudieran presentarse en el futuro.

La creación de escenarios que pudieran pasar bajo circunstancias específicas son una buena ayuda. Generar estos escenarios de forma manual o con herramientas básicas resulta inapropiado debido a la cantidad de variantes que pueden surgir, por ello se utilizan modelos que apoyen en la rapidez de dicha relación y ejecución de la decisión. Además, crear escenarios permite analizar situaciones presentes que impliquen ajustes en la decisión.

¿Cómo podrían los tomadores de decisiones interpretar de una forma sencilla los cambios en sus modelos? Si en un futuro se tienen que generar nuevos escenarios, ¿qué ajustes podría tener el modelo y así hacer una interpretación sencilla?



Uno de los retos más comunes al usar modelos de aprendizaje automático es interpretar el proceso de generación de resultados. Estos modelos pueden ser clasificados de acuerdo con el tratamiento de las variables independientes. Esto permite identificarlos de dos formas:



**Modelos
lineales**



**Modelos
no lineales**



Los modelos lineales utilizan una transformación lineal (suma ponderada) de las variables independientes para describir a la variable objetivo. En el caso simple, donde solamente se tiene una variable independiente y una variable objetivo, la relación entre x y y se representa con una línea recta. Al agregar más variables independientes, se va aumentando la dimensionalidad de esta recta, generando hiperplanos.

La familia de modelos lineales es representada por el algoritmo GLM (*Generalized linear models*). Existen dos GLM más comunes en la práctica:

Regresión lineal

- Es un modelo lineal generalizado que permite resolver un problema de regresión en donde se describe la relación entre x y y como una distribución normal.

Regresión logística

- Es un modelo lineal generalizado que permite resolver un problema de clasificación en donde se describe la relación entre x y y como una distribución de Bernoulli, la cual se utiliza para mostrar la relación entre dos variables aleatorias discretas cuyo resultado único posible se genera entre dos resultados, por ejemplo, "éxito" o "fracaso".

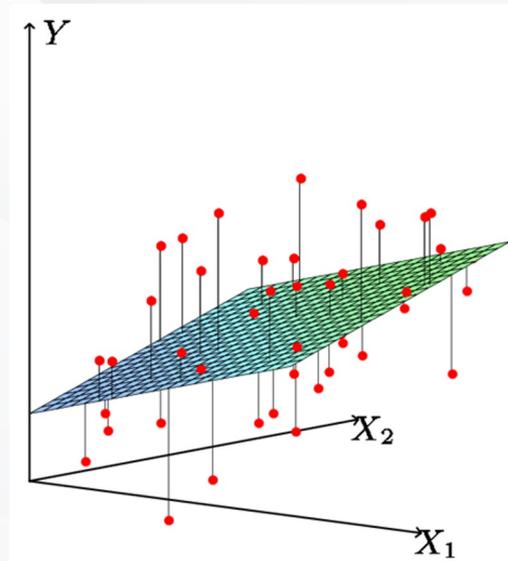


La combinación lineal de las variables independientes X para representar la variable objetivo Y, se representa en la siguiente fórmula:

$$\beta_0 + \beta_1 + \dots + \beta_p \beta_p$$

En donde: B0 es la intersección mientras que B1 es la pendiente.

Así mismo, la representación visual de un plano generado por la combinación lineal de dos variables independientes es la siguiente:





Los modelos lineales permiten describir tendencias en los datos generados por la relación entre las variables independientes y la variable dependiente. La simpleza de los modelos lineales facilita la interpretación del proceso de generación de resultados. Al utilizar una suma ponderada, es decir, una combinación lineal, se pueden tomar directamente las “ponderaciones” para entender cómo y por qué el modelo está generando alguna predicción.

- *Variable independiente numérica.* Por ejemplo, si se quiere estimar el precio final de una casa en función de sus atributos, una posible variable independiente numérica podría ser el número de metros cuadrados del terreno.
- *Variable independiente binaria.* Este tipo de variable generalmente se interpreta como presente (1) o ausente (0). En el ejemplo de la casa, una variable independiente binaria podría ser si la casa tiene o no alberca. La presencia de la categoría asignada a (1) agrega el valor de la ponderación a la estimación final del modelo.
- *Variable independiente categórica (más de 2).* Por ejemplo, considera que una variable de la casa es el tipo de piso, como mármol, madera, etcétera.



Los algoritmos de aprendizaje automático no lineal permiten describir con mejor precisión la variabilidad de los datos usados durante su entrenamiento.

Por otro lado, la composición de los efectos no-lineales convierte virtualmente imposible determinar la contribución que tienen las variables independientes en la construcción del resultado final.

La necesidad de mejorar el rendimiento predictivo de los modelos sin sacrificar la interpretabilidad de los resultados ha dado pie al desarrollo de herramientas sofisticadas en un nicho conocido como *Explainable Machine Learning*. Una de las herramientas más utilizadas se conoce como *SHAP (Shapley Additive Explanations)*.





1. Escoge un modelo de clasificación no lineal disponible en el catálogo de modelos supervisados de scikit-learn (https://scikit-learn.org/stable/supervised_learning.html). Después, utiliza el conjunto de entrenamiento (*train*) para generar el modelo.
 - a. Explica el tipo de modelo seleccionado. Utiliza una investigación propia para brevemente explicar los fundamentos del modelo seleccionado.
 - b. Muestra el código de implementación y desarrollo del modelo con comentarios y explicaciones.
 - c. ¿Cuál es el rendimiento predictivo del modelo de clasificación no lineal? Calcula e interpreta con las métricas propuestas en la introducción.

Si consideras cualquier modelo lineal, ¿el modelo no lineal tiene un mejor o peor rendimiento en el conjunto de prueba respecto al modelos de regresión logística? Explica e interpreta los resultados.



Los modelos de regresión que existen son el lineal y el logístico. El primero permite detallar tendencias con una sola variable, mientras que el segundo lo hace con varias variables. Los modelos no lineales pueden interpretarse utilizando el método SHAP, con el fin de predecir con mayor exactitud las variaciones entre las variables, lo que sin duda generará mayor certeza en la información presentada y con ello tomará mejores decisiones.



SKILLING
CENTER

TECMILENIO



Inteligencia artificial y machine learning

Diseño de proyectos de machine learning y
estrategias de selección de modelos





Los modelos de *machine learning* son muy utilizados como una herramienta de apoyo en la toma de decisiones. En una institución financiera es posible que se hayan generado para la evaluación de solicitudes crediticias y categorización de los clientes en cuanto al nivel de riesgo, liquidez y solvencia a corto, mediano y largo plazo,

Para realizar un análisis de ellos, requiere conocer la contribución que cada uno de los modelos creados tiene que generar. Es decir, debe evaluar la forma en la que la información generada ha contribuido a una decisión efectiva, por ejemplo, si se ha requerido alguna información adicional a considerar, esto es, si existe alguna nueva variable. ¿Cómo se podría detallar esta contribución?, ¿cuál podría ser una estrategia veraz para seleccionar el modelo que se manejará dentro del ERP (planificación de los recursos empresariales), si este fuera el caso? y ¿será conveniente detallar un nuevo modelo para esta evaluación?

Dentro de esta experiencia educativa, se revisarán temas que, sin duda, ayudarán a considerar una solución a la problemática planteada.

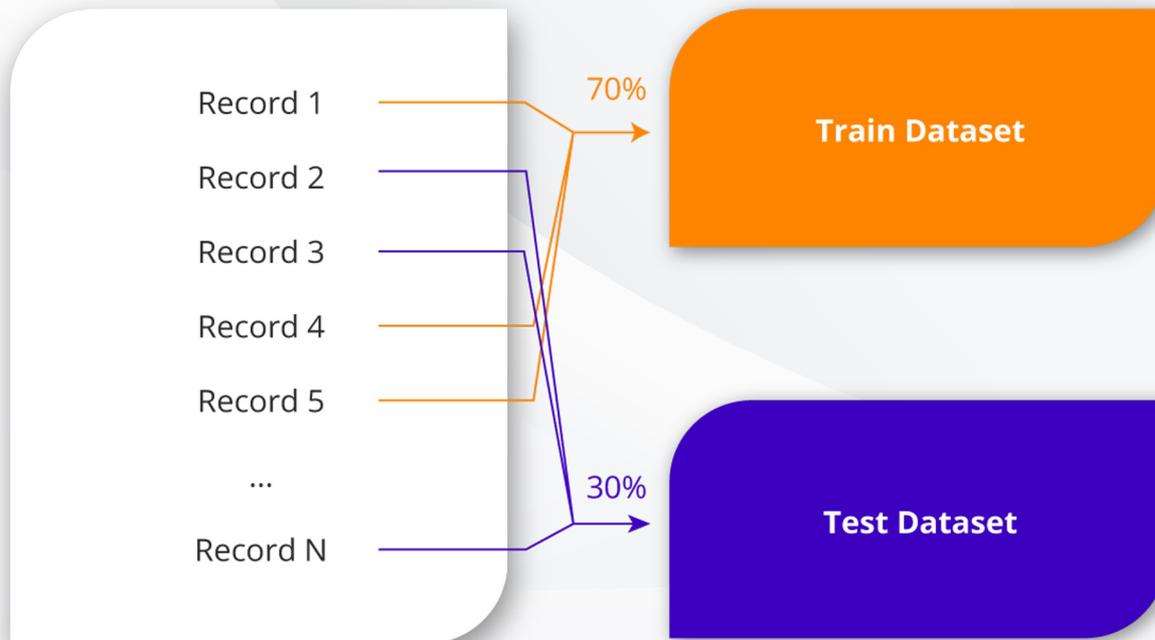


Los algoritmos de aprendizaje automático supervisado utilizan la información de un conjunto de datos para determinar el mejor valor de los parámetros del modelo. El criterio para determinar qué es un “buen” valor de estos parámetros se basa en la capacidad de minimizar una función de error o maximizar una función de ganancia. De esta forma, es posible argumentar que un algoritmo de *machine learning* adquiere la forma y valores que se pueden derivar de los datos. Por consecuencia, el valor de los parámetros modelo es dependiente de la información que se utiliza al momento del entrenamiento. Esta información debe ser representativa del contexto donde el modelo va a operar.



La capacidad del modelo de operar de forma eficiente sobre diferentes conjuntos de datos se conoce como generalización. Una forma para probar la capacidad de generalización del modelo es separar el conjunto de datos en dos subconjuntos: conjunto de entrenamiento (train) y conjunto de prueba (*test*).

Es común que en la práctica se utilice aproximadamente un 70 % de los datos para el entrenamiento.





Hasta ahora se ha identificado que los parámetros que componen un modelo se definen en función de los datos que se utilizan al momento del entrenamiento. En la práctica se tiene la libertad de modificar el valor de los hiperparámetros para escoger aquellos que mejoren el rendimiento predictivo del modelo.

Por lo general se puede dividir el conjunto de entrenamiento original en K partes (subconjuntos) con un número incremental asignado comenzando del 1 hasta K . De esta forma, se utiliza una parte (subconjunto) para validación y el resto (complemento $K-1$) para entrenamiento. Esto se aplica en K -iteraciones, donde se toma el subconjunto i para validación y el complemento para entrenamiento.

Como resultado se tendrán K modelos entrenados de forma independiente bajo la misma configuración y K estimaciones de rendimiento estadístico sobre el conjunto de validación. Usar el promedio del rendimiento estadístico sobre los subconjuntos de validación genera una mejor estimación para comparar diferentes resultados al variar los hiperparámetros. Este algoritmo se conoce como *K-Fold Cross-Validation*. Según Kuhn y Johnson (2018), utilizar K entre 5 y 10 es una práctica común en la industria.



De acuerdo con Li et al. (2020), durante el desarrollo de modelos de aprendizaje automático siempre es recomendable utilizar un modelo de referencia. Este modelo debe ser simple y servir como punto de comparación respecto al modelo que se está desarrollando. En la práctica, se tienen dos casos comunes para el aprendizaje supervisado:

- Caso de regresión en donde la variable objetiva es un valor numérico generalmente continuo.
- Caso de clasificación en donde la variable objetiva es binaria o categórica.

En la industria financiera, los modelos sofisticados de aprendizaje automático pertenecen a la categoría de modelos no lineales, usualmente con base en árboles de decisión o redes neuronales. Estos modelos generalmente tienen mejor capacidad predictiva en comparación con los modelos lineales.





Todos los modelos de *machine learning* tienen parámetros predefinidos, los cuales limitan la búsqueda para elegir la mejor combinación de datos. *Grid search* permite hacer una búsqueda ordenada de la mejor combinación de hiperparámetros. Esto con el fin de limitar el número de combinaciones posibles de acuerdo con los parámetros que se están buscando. Sin embargo, entre más exhaustivo y preciso se quiera ser en la búsqueda, más combinaciones se tienen que ejecutar. Por lo tanto, el consumo de recursos computacionales incrementará y así también el tiempo que conlleva generarlos, lo que puede ser una limitante importante. Por otro lado, *random search* es una forma que apoya la búsqueda generando una búsqueda aleatoria de combinaciones y Bergstra y Bengio (2012) han demostrado que es más eficiente computacionalmente que *grid search*. Considera que, al ser aleatorio, probablemente no se encontrará la combinación perfecta de hiperparámetros, pero se ahorrará en tiempo.



Para la resolución del siguiente problema es necesario utilizar el conjunto de datos disponible en este repositorio en la nube

(<https://gist.github.com/RHDZMOTA/f6c5b6429fc0b51fda2deae2218d04b8/>)

Realiza lo siguiente:

1. Lee la sección de *abstract* de la publicación, *Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study* (<https://www.sciencedirect.com/science/article/abs/pii/S0377221716000412>)
2. Utiliza esta información y complementa con investigaciones propias para generar una introducción explicando el contexto de este problema. Considera estas preguntas:
 - a. ¿Por qué es importante desarrollar modelos de predicción de bancarrota para las instituciones financieras?
 - b. ¿Cuáles son las variables independientes que se suelen utilizar en estos modelos?
 - c. ¿Cuál es el criterio de selección y métricas recomendadas para evaluar el modelo?
3. Realiza un análisis exploratorio de los datos sobre el conjunto de datos completo.
4. Particiona el conjunto de datos en subconjuntos de entrenamiento (80 % *train*) y prueba (20 % *test*).
 - a. ¿Por qué es importante dividir los datos en estos dos conjuntos?



La descarga de la información se puede hacer programáticamente con *Python*:

```
import json
import pandas as pd
import request
url_base = (
    "https://gist.github.com/RHDZMOTA/"
    "f6c5b6429fc0b51fda2deae2218d04b8/raw/"
    "b4fla7a75396e2dfcaeeef1506f7418a061e5531/"
    "company - bankruptcy - records. Json"
)

dataset = pd.DataFrame([json.loads (row) for row in requests.get
(url).text.splitlines()])
print(dataset.shape)
dataset
(1000,96)
```



Ahora que identificas la forma en la que se puede disminuir el sesgo de la información, evaluar el impacto de los modelos para la toma de decisiones será mucho más sencillo. Además, ahora tienes herramientas que te permitirán generar una clasificación apropiada de las variables numéricas para detallar una probabilidad de ocurrencia en las categorías que el área requiere considerar para tomar decisiones efectivas en el otorgamiento de los créditos, amplitudes, o bien, nuevas ofertas para los clientes.

El futuro es incierto y nadie puede saber con exactitud lo que sucederá, sin embargo, es muy probable tener certeza del mismo cuando se tienen las herramientas adecuadas y se cuenta con la información necesaria para sacar el mayor provecho al uso del *machine learning* en el ámbito financiero. Esto te permitirá tener resultados asertivos.



- Bergstra, J., y Bengio, Y. (2012). *Random search for hyper-parameter optimization*. *Journal of Machine Learning Research*, 13. Recuperado de <https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>
- Kuhn, M., y Johnson, K. (2018). *Applied predictive modeling*. Springer.
- Li, D., Hasanaj, E., y Shuo, L. (2020). *3 – Baselines*. Recuperado de <https://blog.ml.cmu.edu/2020/08/31/3-baselines/>

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.

Todos los derechos reservados @ Universidad Tecmilenio

La obra presentada es propiedad de ENSEÑANZA E INVESTIGACIÓN SUPERIOR A.C. (UNIVERSIDAD TECMILENIO), protegida por la Ley Federal de Derecho de Autor; la alteración o deformación de una obra, así como su reproducción, exhibición o ejecución pública sin el consentimiento de su autor y titular de los derechos correspondientes es constitutivo de un delito tipificado en la Ley Federal de Derechos de Autor, así como en las Leyes Internacionales de Derecho de Autor. El uso de imágenes, fragmentos de videos, fragmentos de eventos culturales, programas y demás material que sea objeto de protección de los derechos de autor, es exclusivamente para fines educativos e informativos, y cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por UNIVERSIDAD TECMILENIO. Queda prohibido copiar, reproducir, distribuir, publicar, transmitir, difundir, o en cualquier modo explotar cualquier parte de esta obra sin la autorización previa por escrito de UNIVERSIDAD TECMILENIO. Sin embargo, usted podrá bajar material a su computadora personal para uso exclusivamente personal o educacional y no comercial limitado a una copia por página. No se podrá remover o alterar de la copia ninguna leyenda de Derechos de Autor o la que manifieste la autoría del material.