Los datos utilizados para este informe se pueden encontrar en: Link.

Introducción

En este documento se describirán los resultados obtenidos tras realizar una serie de pruebas sobre el cálculo de los Shapley Values, una herramienta utilizada en XAI (Explainable AI).

Explicación de los Shapley Values

Los Shapley values son una técnica post-hoc que se basa en teoría de juegos. Su objetivo es obtener la contribución marginal esperada de cada jugador (cada variable en nuestro caso).

Para ello se deben considerar todas las posibles combinaciones de todos los posibles valores de las variables en nuestro modelo, cosa que computacionalmente no es beneficiosa. En caso de modelos de un tamaño considerable el tiempo de cómputo puede ser alto (especialmente si no se trata de un modelo basado en árboles), aunque se podría realizar el cálculo sobre un subconjunto de nuestros datos para agilizar este proceso (a coste de perder precisión).

Para más información sobre el cálculo de esta métrica se pueden consultar los siguientes recursos así como la librería de shapley values más usada SHAP**.**

Realización del experimento

Para realizar una prueba sobre estos se han ejecutado cinco modelos sobre una dataset de concesión de crédito. Los modelos seleccionados han sido: decisión tree, multilayer perceptron, random forest, XGBoost y K Nearest Neighbour.

Finalmente, una vez obtenidos todos los shapley values medios para cada posible valor de nuestro modelo se analizan en el notebook adjunto al github del experimento. En el apartado de métricas se explicará de manera más extendida dicho análisis.

Problemas al realizar el experimento

El principal problema encontrado en el experimento fue que los datos seleccionados eran insuficientes y estaban altamente desbalanceados. Debido a esto, los shapley values no son comparables entre modelos para ver el impacto de las variables ya que, al no tener modelos precisos, estos no generalizan bien y parte de sus pesos se deben a carencias de entrenamiento del propio algoritmo.

Por otra parte, remarcar que la librería SHAP no cuenta con una documentación de calidad. Por este motivo encuentran muchas funcionalidades y muchos resultados no documentados o documentos pobremente cosa que dificulta en gran medida el uso de esta herramienta.

Finalmente remarcar que la librería SHAP ha presentado varios errores y resultados no coherentes durante el desarrollo del experimento. Para algún tipo de modelos genera solo un listado de Shapley values mientras que para otros genera un listado para el impacto en las predicciones positivas y otro para las negativas. También se han encontrado errores en el cálculo de los shapley values, aunque en este caso la propia librería los detecta y avisa al usuario.

Explicabilidad de un modelo

A pesar de los problemas descritos, los datos han servido para ganar más conocimiento sobre el porqué un algoritmo tomó una decisión u otra mirando sus shapley values y gracias a esto si que se han podido detectar sesgos muy marcados en nuestros datos (no en nuestro modelo). Como por ejemplo observar que los datos no son representativos de algunos colectivos.

En el caso de nuestro dataset particular encontrábamos que las personas con estudios superiores eran mucho más propensas a ser clientes de riesgo, pero al observar la cantidad de este tipo de clientes observamos que no era un conjunto representativo (menos de 20).

El problema es que estos valores solo han servido para sesgos muy marcados como el descrito anteriormente, que probablemente se podrían eliminar o reducir al hacer un buen EDA y limpiado de datos. En el siguiente apartado, donde se analiza la comparación entre modelos, se ve que los Shapley values calculados con la librería SHAP parecen no ser tan efectivos para observar el impacto de todas las variables.

Comparación de modelos

Los datos obtenidos se muestran al final de este apartado.

Para realizar la comparación de los modelos se ha calculado una media de sus Shapley Values y se han normalizado sus valores entre -100 y 100.

Para medir el grado de similitud entre los valores medios de cada modelo, se han utilizado las métricas siguientes:

- R2 score
- Mean squared error (MSE)
- Mean absolute error (MAE)

Como se puede ver en los datos adjuntos en la parte final de este apartado, la comparación entre modelos es bastante negativa.

El R2 score al ser prácticamente siempre negativo o cercano a 0 nos indica que nuestros datos tienen poca similitud.

Por otra parte, el MSE nos muestra valores extremadamente altos y el MAE nos muestra valores más bajos. Con esto deducimos que aunque haya variables con pesos similares encontramos otras con valores muy dispares, cosa no aceptable ya que nos indica que las variables de alto impacto son diferentes en los shapley values de cada modelo.

Tras ver estos resultados, se probó de realizar el experimento sobre otro dataset mejor balanceado (Titanic dataset) usando RF y XGBoost. A pesar que los valores mejoraron, siguieron sin ser aceptables.

conclusiones de comparación de modelos

Tras estos resultados, si se consideran válidos los métodos utilizados en el experimento, podríamos concluir que los SHAP values no son efectivos para comparar diferentes modelos.

Esto podría ser debido a que la librería SHAP calcula una aproximación de los Shapley Values y no su valor real.

Aunque para afirmar rotundamente este hecho, se deberían de realizar más pruebas sobre otros conjuntos de datos.

Datos obtenido

Prueba sobre dataset de crédito

Datos obtenidos con las métricas:

r2 Score for xgboost and DT : -2.725319324844746
Mean squared error for xgboost and DT : 198.08149938453263
Mean absolute error for xgboost and DT : 6.782446089571855

r2 Score for xgboost and MLP : -0.8146972934848373
Mean squared error for xgboost and MLP : 96.49050980012576
Mean absolute error for xgboost and MLP : 5.3778301126089465

r2 Score for xgboost and KNN : 0.05387685053764124
Mean squared error for xgboost and KNN : 50.30696047934915
Mean absolute error for xgboost and KNN : 3.7994892812769656

r2 Score for DT and xgboost : -1.8438338302959645
Mean squared error for DT and xgboost : 198.08149938453263
Mean absolute error for DT and xgboost : 6.782446089571855

r2 Score for DT and MLP : -2.3594482468343965
Mean squared error for DT and MLP : 233.99557975173337
Mean absolute error for DT and MLP : 6.28425891312816

r2 Score for DT and KNN : -1.633271569049478
Mean squared error for DT and KNN : 183.41521052574902
Mean absolute error for DT and KNN : 4.973010457793383

r2 Score for MLP and xgboost : -0.5753370913754321
Mean squared error for MLP and xgboost : 96.49050980012576
Mean absolute error for MLP and xgboost : 5.3778301126089465

r2 Score for MLP and DT : -2.820291930930633
Mean squared error for MLP and DT : 233.99557975173337
Mean absolute error for MLP and DT : 6.28425891312816

r2 Score for MLP and KNN : -0.9531629027141424
Mean squared error for MLP and KNN : 119.63260767321519
Mean absolute error for MLP and KNN : 5.845758918260783

r2 Score for KNN and xgboost : 0.2424257543843531
Mean squared error for KNN and xgboost : 50.30696047934915
Mean absolute error for KNN and xgboost : 3.7994892812769656

r2 Score for KNN and DT : -1.7620559545734866
Mean squared error for KNN and DT : 183.41521052574902
Mean absolute error for KNN and DT : 4.973010457793383

r2 Score for KNN and MLP : -0.8015515476486048
Mean squared error for KNN and MLP : 119.63260767321519
Mean absolute error for KNN and MLP : 5.845758918260783

Prueba sobre Titanic df

r2 Score for XGBoost and Random Forest : -0.7113408837755708
Mean squared error for XGBoost and Random Forest : 20.297307542526536
Mean absolute error for XGBoost and Random Forest : 0.9431450374092838

r2 Score for XGBoost and Random Forest : -0.7161587159293241
Mean squared error for XGBoost and Random Forest : 20.297307542526536
Mean absolute error for XGBoost and Random Forest : 0.9431450374092838

Reconocimiento de los Shapley values

A pesar de los datos obtenidos, parece ser que el uso de la librería Shap ha crecido de manera notable en el último periodo. Esto se puede observar debido a la creciente publicación de artículos, conferencias y posts que se observa en la red. Por otra parte, la plataforma Kaggle (una plataforma de prestigio relacionada con IA) ha incluido a la librería SHAP como parte de formación en nuevo módulo enfocado a la XAI.

Conclusiones Finales

Según los datos y procedimientos seguidos en el experimento, mediante el uso de la librería SHAP **se pueden detectar sesgos muy marcados en nuestro modelo o nuestros datos** analizando outliders en los Shapley Values. Sin embargo, **para valores que aportan una contribución menor la librería SHAP no es lo suficientemente precisa** al calcular una aproximación de los shapley values,

Por otra parte, a la hora de comparar modelos o evaluar el peso de las variables dentro de rangos normales para la predicción de un valor se ha encontrado que los datos no son consistentes. **Esto podría ser debido a que la librería SHAP no calcula los Shapley Values reales si no que realiza una aproximación de estos suponiendo que no existe una correlación entre las variables, cosa que rara vez se da en la práctica.**

Sin embargo, el crecimiento de las menciones a los Shapley Values en la red podría indicar que los métodos utilizados en el experimento no son adecuados o que se deberían de estudiar otros datos y/o otros modelos para obtener unas conclusiones correctas.

También es importante destacar que actualmente la librería SHAP no presenta una documentación adecuada ni una estructura consistente en todos sus métodos y clases.

En conclusión, a pesar que los Shapley Values teóricamente son una herramienta muy potente, parece ser que en la práctica al utilizar la librería SHAP no tienen un uso real que pueda aportar explicabilidad a nuestro modelo (probablemente debido a la forma en que SHAP realiza en cálculo de estos valores). Sin embargo, el creciente uso de la librería podría indicar que se necesita de más experimentación con esta para poder afirmar estas conclusiones.

Los datos y el procedimiento utilizados para este informe se pueden encontrar en: Link.