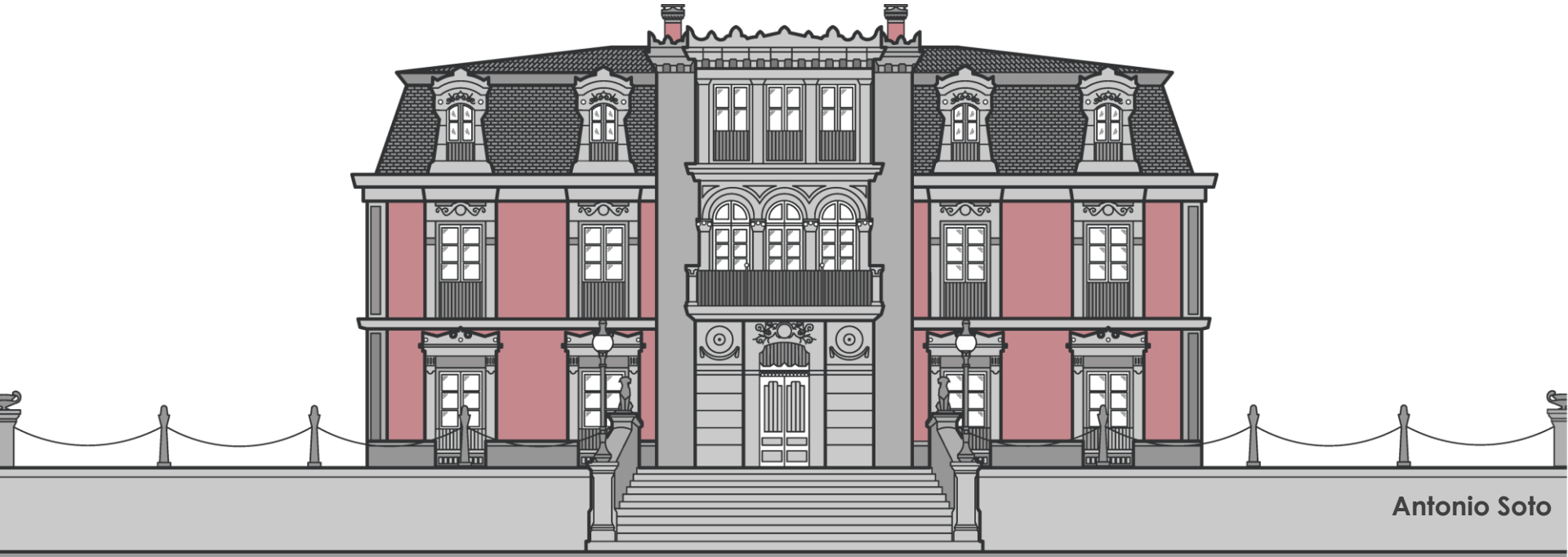
IFFE Business School

Interpretando Modelos de Machine Learning



Antonio Soto

CEO SolidQ

31 de Mayo de 2019

MBD&BA

¡HOLA!

CEO

SolidQ

Soy Antonio Soto, llevo más de 20 años dedicado al mundo del análisis de datos, desde los primeros sistemas Data Warehouse, hasta el desarrollo de soluciones integradas con Inteligencia Artificial de hoy en día, pasando por todos los puntos intermedios. Me he tocado viajar por el mundo diseñando sistemas y soluciones enfocadas a la toma de decisiones en todos los sectores y en empresas de diferentes tamaños.

Sígueme en…

asoto@solidq.com

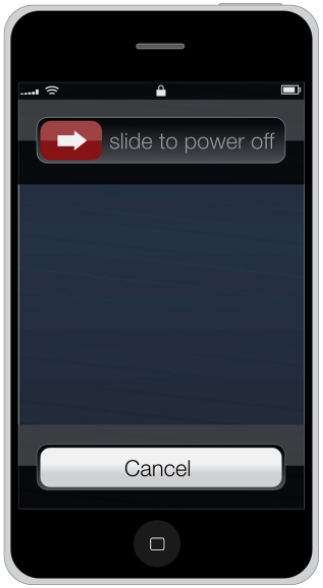
+34.637.505.941

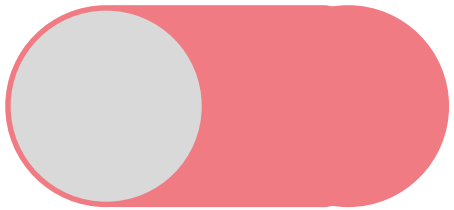
https://www.linkedin.com/in/antoniosql

@antoniosql

MBD&BA

RECUERDA:





APAGA EL MÓVIL

No lo olvides!!

¿Qué aprenderemos?

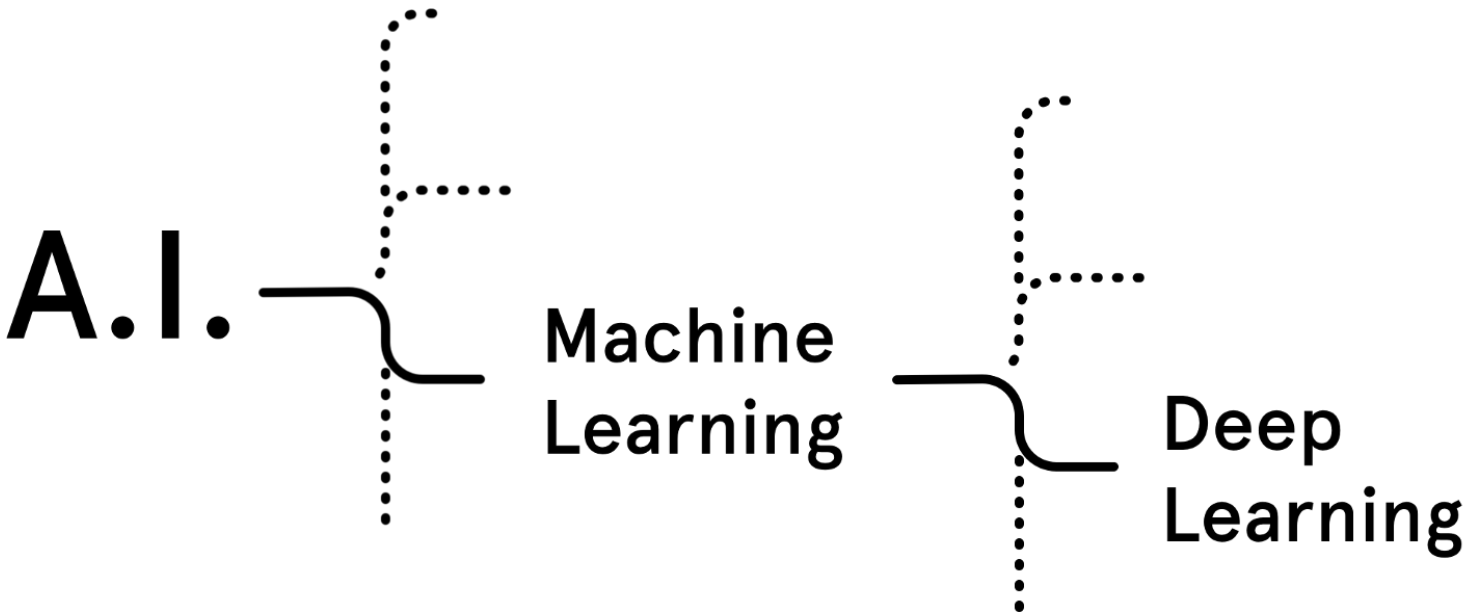
1. Introducción a Machine Learning
2. MOTIVACIONES: Importancia de la interpretación de modelos ML
    1. Aspectos legales
    2. Aspectos éticos
    3. Aspectos de negocio
3. TEORIA DE LA INTERPRETACIÓN DE MODELOS
4. Estrategias de Interpretación de Modelos
    1. Técnicas tradicionales
    2. Limitaciones de técnicas tradicionales
    3. Técnicas ACTUALES de Interpretación

IFFE Tech School

INTRODUCCIÓN A MACHINE LEARNING

IFFE Tech School

MBD&BA


A.I.
Machine Learning
Deep Learning

¿Qué es Machine Learning?

Proceso de utilizar datos para tomar decisiones

Determinar que tipo de algoritmos utilizar basándose en los datos

- ¿Tenemos un conjunto finito de respuestas?
- ¿Estamos buscando respuestas de algún tipo basadas en ciertos datos?

Debemos de definir claramente el problema

- Debe de responder a una pregunta específica

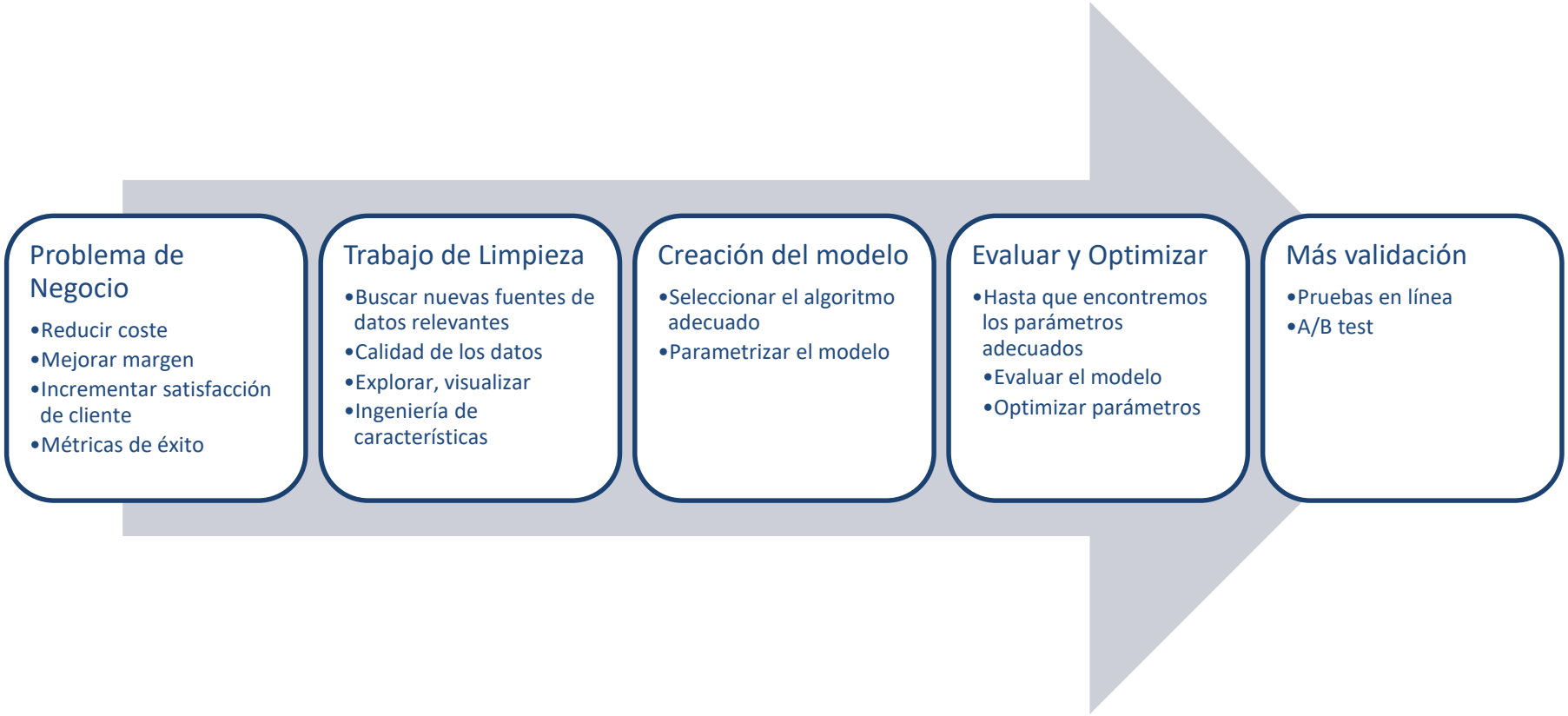
Tipos de ML

- Aprendizaje Supervisado tiene un conjunto definido de entradas y salidas
  - Datos etiquetados
  - Feedback directo
  - Predice salida / futuro
- Aprendizaje No Supervisado tiene entradas pero las salidas son desconocidas
  - No tenemos etiquetas
  - No feedback
  - Busca estructuras ocultas en los datos
- Aprendizaje Reforzado
  - Proceso de decisión
  - Sistema de recompensa
  - Aprende una serie de acciones

Antes de nada…entiende los datos

- ¿Qué pregunta estoy intentando resolver?
- ¿Cuál es el mejor modo de expresar la pregunta como un problema de Machine Learning?
  - ¿Podemos adivinar si un cliente nos va a abandonar?
- ¿Disponemos de datos detallados o tenemos resúmenes?
- ¿Cómo mediremos el éxito?
- ¿Cómo interactuará con otros componentes?

Ciclo de vida de un proyecto ML

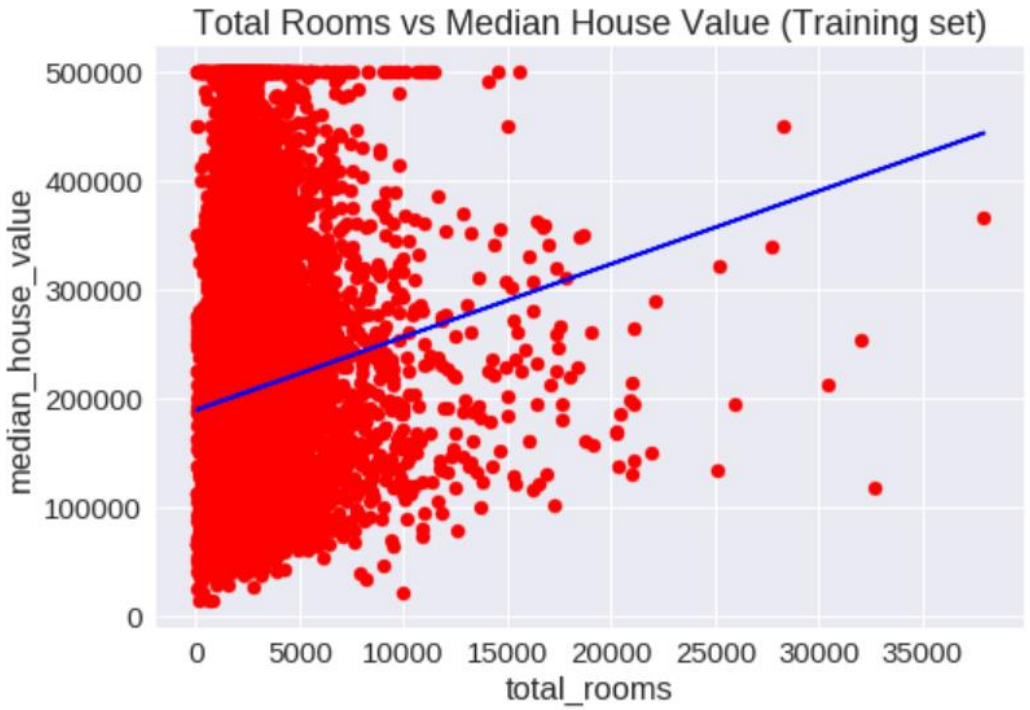


Demostración Algoritmos de Clasificación

- Algoritmos para Clasificación
  - Árboles de Decisión
  - k-NN
  - Regresión Logística
  - SVM
- Fiabilidad de los modelos
- Matriz de Confusión

Demostración Algoritmos de Regresión

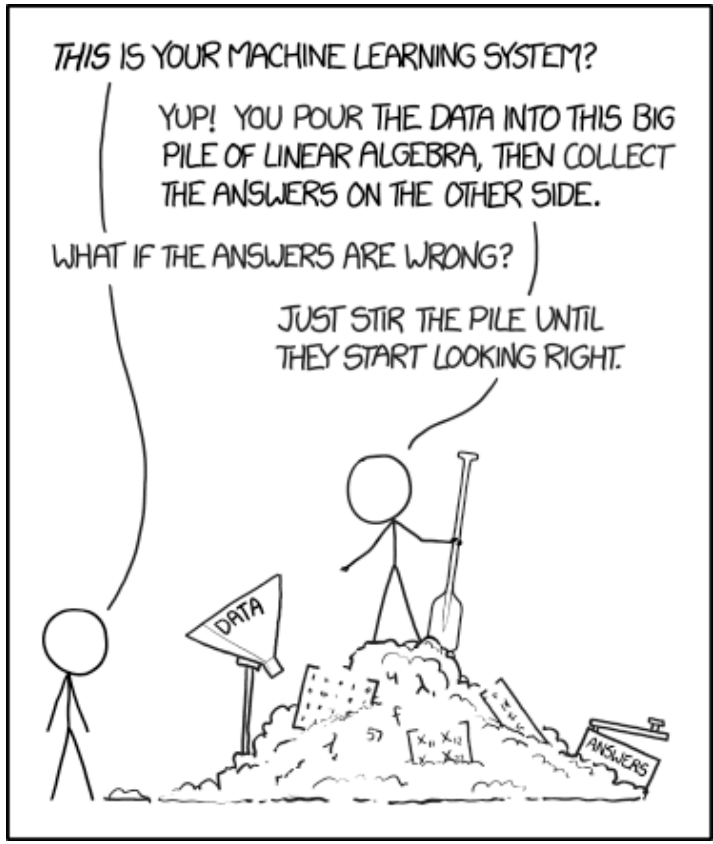
- Regresión Lineal
- Regresor k-NN



MOTIVACIONES

La importancia del ML Interpretable

- ¿Cómo se que puedo confiar en el modelo?
- ¿Cómo toma sus decisiones?
- Balance entre rendimiento e interpretabilidad
- Interpretaciones
  - Global ¿Cómo hace las predicciones? ¿Cómo influyen los subconjuntos de datos?
  - Local ¿Por qué ha tomado una decisión en concreto para un caso en concreto?



Motivaciones: importancia de la interpretación de Modelos de Machine Learning (I)

- Legales
  - GDPR y Derecho a la explicación
    - Artículos 13-15
      - Derecho a entender como se utilizan sus datos
      - Decisiones automáticas → "Información significativa sobre la lógica involucrada, importancia y consecuencias de la decisión"
    - Artículos 21-22
      - Información suficiente para excluirse del proceso
    - Recital 71
      - "derecho a obtener la intervención humana, para expresar su punto de vista, para obtener una explicación de la decisión alcanzada después de dicha evaluación y para desafiar la decisión"
  - Información de base:
    - Información técnicas
      - Algoritmo, lógica del Algoritmo, de donde provienen los datos y cuantas características se están utilizando
    - Información de despliegue
      - Cuándo, cómo y para qué se utiliza
    - Educar al sujeto

Motivaciones: importancia de la interpretación de Modelos de Machine Learning (y II)

- Éticas
  - Sesgo
- De Negocio
- Beneficios
  - Dar confiabilidad en los resultados.
  - Ayudar en la Depuración.
  - Informar a la Ingeniería de Características (Feature Engineer).
  - Detectar necesidad de recoger nuevas muestras.
  - Ayudar a una persona en la toma de decisiones.
  - Mayor seguridad/robustez en el modelo obtenido.

TEORÍA DE LA INTERPRETACIÓN DE MODELOS

Entendiendo la Interpretación de un modelo

- Intentar comprender y explicar las decisiones tomadas por un modelo al realizar sus predicciones

  - ✓ ¿Qué conceptos dirigen las decisiones del modelo?
    - **JUSTICIA**
  - ✓ ¿Por qué el modelo ha tomado una determinada decisión?
    - **RESPONSABILIDAD**
  - ✓ ¿Cómo podemos confiar en las predicciones del modelo?
    - **TRANSPARENCIA**

Criterios para seleccionar un método

- Analizando los resultados del método
  - ✓ Estadísticas de resumen de las Características
  - ✓ Visualización del resumen de Características
  - ✓ Información interna del modelo
  - ✓ Punto de Datos
  - ✓ Modelo intrínsecamente interpretable
- Específico para un modelo o agnóstico al modelo
- Local o Global

Ámbitos de la Interpretabilidad

- Transparencia del Algoritmo
  - ✓ ¿Cómo el algoritmo crear el modelo?
- Interpretabilidad Holística del Modelo, Global
  - ✓ ¿Cómo hace las predicciones el modelo entrenado?
- Interpretabilidad Global del Modelo en un Nivel Modular
  - ✓ ¿Cómo afectan partes del modelo a sus predicciones?
- Interpretabilidad Local para una única predicción
  - ✓ ¿Por qué hace una determinada predicción el modelo para una instancia concreta de datos de entrada?
- Interpretabilidad Local para un Grupo de predicciones
  - ✓ ¿Por qué el modelo hace determinadas predicciones para un grupo de instancias?

Las ***Explicaciones*** *(I)*

Una explicación relaciona los valores de las características de una instancia, con sus predicciones de un modo entendible por las personas.

- Propiedades del Método de Explicación
  - ✓ Poder de expresividad
  - ✓ Translucidez
  - ✓ Portabilidad
  - ✓ Complejidad algorítmica

Las ***Explicaciones*** *(y II)*

- Propiedades de las explicaciones individuales
  - ✓ Precisión
    - ¿Cómo de bien predice una explicación los datos no vistos?
  - ✓ Fidelidad
    - ¿Cómo de bien se aproxima la explicación de la predicción del modelo de caja negra?
  - ✓ Consistencia
    - ¿Cuánto difiera una explicación entre los modelos que se han entrenado en la misma tarea y que producen predicciones similares?
  - ✓ Estabilidad
    - ¿Cómo de iguales son las explicaciones para casos similares?
  - ✓ Comprensión
    - ´¿Cómo de bien entienden las explicaciones las personas?
  - ✓ Certeza
    - ¿Refleja la explicación la certeza del modelo de ML?
  - ✓ Grado de importancia
    - ¿Cómo de bien se reflejan la importancia de las características?
  - ✓ Novedad
    - ¿Refleja si una instancia de datos que se va a explicar proviene de una región "nueva" alejada de la distribución de datos de entrenamiento?
  - ✓ Representatividad
    - ¿Cuántos casos cubre una explicación?

ESTRATEGIAS DE INTERPRETACIÓN DE MODELOS

Técnicas Tradicionales

- Técnicas EDA
  - ✓ Reducción de Dimensionalidad (PCA)
- Métricas de Evaluación de rendimiento de nuestro modelo
  - ✓ Aprendizaje Supervisado – Clasificación
    - Matriz de Confusión
    - Accuracy, Precision, Recall, F1-Score
    - ROC y AUOC score
  - ✓ Aprendizaje Supervisado – Regresión
    - Coeficiente de determinación (R-square)
    - RMSE (root mean-square error)
    - MAE (Mean absolute error)

Limitaciones de Técnicas Tradicionales

- Medimos el rendimiento
- ¿Cumple con parámetros de rendimiento válidos?
- Con estas técnicas, ¿cómo aseguramos?
  - ✓ JUSTICIA
  - ✓ RESPONSABILIDAD
  - ✓ TRANSPARENCIA
- Balance entre Rendimiento e Interpretabilidad

Demostración Técnicas Tradicionales

- Demostración EDA

- Demostración Importancia Escalado y PCA
  - Concepto de Escalado
  - PCA y análisis EDA

Técnicas Adicionales de Interpretación de Modelos

1. Uso de Modelos interpretables
2. Importancia de Características
3. Plots de dependencia parcial (PDP)
4. Modelos Subrrogados Globales
5. LIME (Local Interpretable Model-agnostic Explanations)
6. SHAP (Shapley Values y SHapley Additive exPlanations)

Uso de Modelos Interpretables

- Cómo categorizarlos según sus características
  - Linealidad
  - Monotonicidad
  - Interacciones

Algoritmo	Lineal	Monótono	Interacción	Tarea
Modelos Lineales	SI	SI	NO	Regresión
Regresión Logística	NO	SI	NO	Clasificación
Árboles de Decisión	NO	Algo	SI	Clas + Regre
Naive Bayes	SI	SI	NO	Clasificación
K-NN	NO	NO	NO	Clas + Regre

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Importancia de Características

- Importancia de una Característica
  - Incremento en el error de predicción del modelo después de permutar los valores de la característica, lo que rompe la relación entre la característica y la salida
- ¿Cómo lo calculamos?
  - Sobre datos de entrenamiento
  - Sobre datos de test
- Ventajas
  - Interpretación sencilla
  - Visión global comprimida del modelo
  - Las mediciones entre medidas son comparables
  - No requiere re-entrenar el modelo
- Desventajas
  - No está claro si debemos usar conjunto de entrenamiento o de prueba
  - Está ligado al Rendimiento (error) no a como cambia la salida
  - Necesitamos la salida real
  - Si existe mucha correlación entre características, repartirá la importancia entre ellas

Demostración

Importancia de Características

- El propio modelo
- Skater

Partial Dependence Plot (PDP)

- PDP muestra el efecto marginal que una o dos características tienen en una predicción
- Muestra si la relación entre una característica y la salida es lineal, monótona o más compleja
- Ventajas
  - Intuitivo
  - Fácil de implementar
  - Interpretación es clara
- Desventajas
  - EL número máximo de entidades es dos
  - Asume independencia de las características
  - Muestra medias de efectos marginales, por lo que no veremos extremos

Demostración Dependencias Parciales

- PDP
- Uso de SKATER para PDP

Modelos de Sustitución Globales (I)

- Modelo interpretable entrenado para aproximar la salida de un Modelo Caja Negra
- Pasos
  - Seleccionar un conjunto de datos X
  - Obtener las predicciones del Modelo Caja Negra para ese conjunto de datos X
  - Seleccionar un modelo interpretable (regresión lineal, árbol de decisión,…)
  - Entrenar el modelo interpretable con el conjunto de datos X y obtener sus predicciones
  - YA TENEMOS EL MODELO de SUSTITUCIÓN
  - Medir como de bien replica las predicciones del Modelo Caja Negra
  - Interpretar el modelo de sustitución

Modelos de Sustitución Globales (y II)

- Ventajas
  - Flexible
  - Intuitivo
  - Con la medida de R-square podemos saber fácilmente como aproxima un modelo a otro
- Desventajas
  - Estamos sacando conclusiones sobre el modelo, no sobre los datos
  - No hay una medida de clara que valor de relación de R-square es bueno
  - Si estamos utilizando un subconjunto, estamos aproximando para ese subconjunto

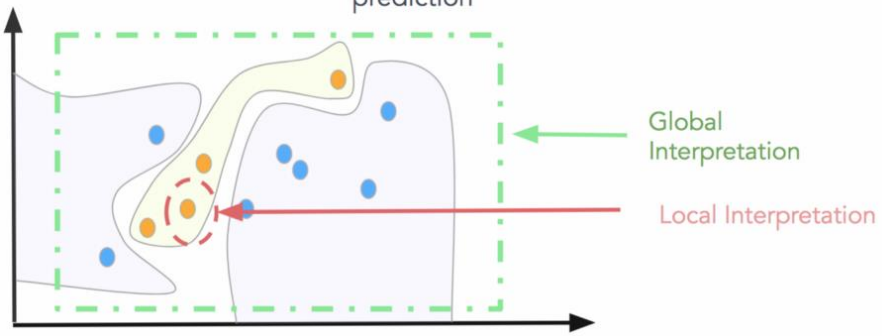
Modelos de Sustitución con Skater

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



LIME (Local Interpretable Model-agnostic Explanations) (I)

- Explicaciones para una predicción individual de un modelo de caja negra
- Concepto
  - Nos olvidamos de datos de entrenamiento
  - Tenemos una caja negra a la que podemos pasarle las instancias que queramos las veces que queramos para ver que predicciones genera
  - Objetivo: Entender porque hace cada una de esas predicciones
  - LIME prueba que ocurre con las predicciones cuando se varían los datos de entrada
  - Genera un conjunto de datos compuesto de muestras permutadas y su correspondiente predicción por parte del modelo de Caja Negra
  - Entrena un modelo interpretable con ese conjunto de datos

LIME (Local Interpretable Model-agnostic Explanations) (y II)

- Ventajas
  - Incluso si cambiamos el modelo de caja negra podemos seguir utilizando el modelo interpretable
  - Las explicaciones son cortas y pueden contrastarse
  - Nos vale para datos tabulados, textos e imágenes
  - Medida de fidelidad
  - Muy fácil de utilizar y con paquetes en R y Python
  - Podemos utilizar características diferentes a las del modelo original
- Desventajas
  - La correcta definición de como de "parecidos" son los modelos es compleja
  - Los datos se obtiene de una distribución Gaussiana que no tiene en cuenta correlación entre características
  - Muy prometedor… pero todavía en desarrollo

Demostración Interpretando Modelos con LIME

- LIME para texto
- Visualización de Explicaciones

SHAP (Shapley Values y SHapley Additive exPlanations) (I)

- Aplicación de Teoría de Juegos
- Intenta explicar una predicción asumiendo que cada característica es un "jugador" en un juego donde la predicción es el "premio"
  - Nos indica como distribuir el "premio" entre los "jugadores"
  - Juego→ Tarea de predicción para una instancia
  - Ganancia→ predicción real menos la predicción promedio para todas las instancias
  - Jugadores→ Valores de las características de la instancia
- Valor de Shapley
  - Contribución marginal promedio de un valor de característica entre todas las posibles relaciones

SHAP (Shapley Values y SHapley Additive exPlanations) (y II)

- De forma "intuitiva"
  - Los valores de las características entran en una habitación en orden aleatorio
  - Todos los valores en la habitación participan en el "juego" (contribuyen a la predicción)
  - El valor de Shapley del valor de una característica es la media de cambio en la predicción que los valores ya existentes en la habitación reciben cuando el nuevo valor de característica se une a ellos
- Ventajas
  - La diferencia entre la predicción y la media de las predicciones se distribuye equitativamente entre los valores de las características de la instancia
    - Propiedad de Eficiencia de los valores de Shapley
  - Teoría sólida
  - Explica la predicción como un juego
  - Contrastable
- Desventajas
  - Mucho tiempo de cómputo
  - Devuelve un único valor por característica no un modelo
  - Necesitamos acceso a los datos si queremos calcularlo para una nueva instancia

Demostración Interpretando modelos con SHAP

- Obteniendo valores SHAP
- Importancia de Características
- PDP

CONCLUSIONES

Conclusiones

1. Examina muy bien el perfil de usuario final
2. Empápate del conocimiento que puedas extraer de los datos
3. Gánate la confianza del usuario a través de interpretaciones y explicaciones que pueda entender

¡Gracias!

Búscame en…

asoto@solidq.com

+34.637.505.941

https://www.linkedin.com/in/antoniosql

@antoniosql

MBD&BA