TP3 - Machine Learning I

Dataset a usar

El dataset a usar es el de la competencia en: https://www.kaggle.com/c/ieee-fraud-detection/
El objetivo del TP **NO** es participar de la competencia y el score logrado **NO** influye en la nota, prestar atención al criterio de corrección y a las consignas.
Los datos están explicados en el siguiente hilo.
Cómo score tanto para validar nosotros como para la competencia vamos a usar AUC-ROC.

Parte I: Análisis exploratorio (6 puntos)

Realizar 6 visualizaciones interesantes que **ayuden a explicar el target**, alguno que contenga al menos un feature de *train_identity.csv* y haciendo al menos un plot de cada tipo:

- Bar plot
- Violin plot
- Box plot
- Heatmap

Parte II: Machine Learning Baseline (4 puntos)

Vamos a construir un modelo muy sencillo para saber qué es lo peor que podemos hacer, en general esta es una tarea muy importante que queremos que repitan en sus proyectos de machine learning. ¿Por qué?

- Navaja de Ockam: “Cuando se ofrecen dos o más explicaciones de un fenómeno, es preferible la explicación completa más simple; es decir, no deben multiplicarse las entidades sin necesidad.” ¿Para qué desarrollar un modelo super complejo si capaz es peor o casi igual que uno muy sencillo?
- Nos sirve para saber si estamos usando bien los modelos más complejos, si su score nos da peor al baseline probablemente se deba a un error de código.
- Nos sirve para rápidamente saber que tan complejo es un problema.
- Los modelos simples son fáciles de entender.

Utilice **todos las columnas del dataset** (exceptuando ids únicos) con algún encoding donde sea necesario para entrenar una regresión logística y utilizando búsqueda de hiperparametros y garantizando la reproducibilidad de los resultados cuando el notebook corriera varias veces. Conteste las preguntas:

- ¿Cuál es el mejor score de validación obtenido? (¿Cómo conviene obtener el dataset para validar?)
- Al predecir con este modelo para la competencia, ¿Cúal es el score obtenido? (guardar el csv con predicciones para entregarlo después)
- ¿Qué features son los más importantes para predecir con el mejor modelo? Graficar.

Parte III: Machine Learning (10 puntos)

Entrenar 2 (de tipos distintos, excluyendo regresiones logísticas) modelos (5 puntos cada uno) con búsqueda de hiperparametros (¿cómo conviene elegir los datos de validación respecto de los de train?).
Los modelos deben cumplir las siguientes condiciones:

- Deben utilizar AUC-ROC como métrica de validación.
- Deben medirse solo en validación, no contra la competencia.
- Deben ser reproducibles (correr el notebook varias veces no afecta al resultado).
- Deben tener un score en validación superior a 0,8.
- Para el feature engineering debe utilizarse imputación de nulos, mean encoding y one hot encoding al menos una vez cada uno.
- Deben utilizar al menos 80 features (contando cómo features columnas con números, pueden venir varios de la misma variable).
- Deben utilizar las columnas: id_31, id_33, DeviceType, DeviceInfo.
- Deben utilizar CountVectorizer o TfIdfVectorizer para algún feature.

Deberán contestar la siguiente pregunta:

- Para **el mejor modelo de ambos**, ¿cuál es el score en la competencia? (guardar el csv con predicciones para entregarlo después)

Puntos extra (hasta 5)

Estas consignas suman puntos extra por fuera de los necesarios para aprobar el TP, mientras más consignas extra realicen más puntos consiguen y menos va a depender su aprobación de que los puntos de arriba estén bien:

- Entrenar una red neuronal con Keras que sea reproducible, usando al menos 80 features, imputación de nulos, mean encoding, one hot encoding y un score en validación superior a 0,7. Debe ser un modelo por separado a los propuestos, no necesita búsqueda de hiper parámetros ni cumplir otra condición. ¿Cúal es su score en validación y en la competencia? (2 puntos)
- Encontrar una relación interesante entre TimestampDT y el target con algún plot (½ punto)
- Graficar la importancia de features para algún modelo de la parte III. ¿Qué tanto se parece a los features importantes de la parte II? (1 punto)
- Graficar la matriz de confusión para algún modelo de la parte III (½ punto)
- Agregar una técnica de feature selection para algún modelo de la parte III (1 punto)

Criterio de corrección

Se necesita un 60% (12/20) de los puntos para aprobar. Los puntos extra permiten sumar por dentro de los 20 (uno se puede sacar hasta 25 pero se sigue aprobando con 12).

Parte I

1. Cada visualización vale un punto, y debe cumplir con las siguientes condiciones:
    a. Debe explicarse por sí misma, sin necesidad de texto aclaratorio.
    b. Debe tener rótulos en los ejes que corresponda y en el título (incluyendo unidades si corresponde).
    c. Debe mostrar una relación con el target que sea clara.
    d. El uso del color debe ser intencional, elegido por ustedes, no por la librería.
    e. La visualización debe ser legible (Un bar chart de 40 barras por ejemplo es ilegible)
    f. Debe cumplir el objetivo propuesto

Parte II

Vamos a corregir los siguientes puntos (no pueden restar más de 4 en total):

- Utiliza mal los datos de validación ya sea para obtener el resultado o para buscar hiper parámetros (-4 puntos), ejemplos: calcular el score con otras labels, calcular el AUC-ROC usando la predicción binaria y no la probabilidad, el set de validación se usa para elegir los parámetros pero también está dentro del entrenamiento de cada modelo, etc.
- El modelo no está bien hecho (-4 puntos), ejemplo: entrenan con las labels o datos cambiados para algunas filas
- No es capaz de predecir para la competencia o no lo hace correctamente (-4 puntos)
- No es reproducible (-2 puntos)
- No obtiene bien los features más importantes (-2 puntos)
- La predicción en la competencia da menos de 0.5 (-2 puntos)
- La predicción para la competencia tiene errores (-1 punto)
- No utiliza todos los features (-1 punto)

Parte III

Vamos a corregir los siguientes puntos en cada modelo de 5 puntos (a medida se acumulan estos pueden hacer que el modelo valga 0, pero nunca negativo):

- Para cada modelo cada condición no cumplida (o mal hecha) resta 1 punto.
- Feature engineering inapropiado para el modelo elegido (-2 puntos), ejemplos: features que no están normalizadas para una red neuronal, features sin ninguna consideración de escalas para un KNN, etc.
- No buscan para todos los hiperparametros importantes.

Además si un modelo **diera un resultado menor a 0,6 en validación** se invalida entero. Por sobre el puntaje total del ejercicio (ambos modelos) se restan 3 puntos si cualquiera de las siguientes cosas suceden (no acumulables): eligen mal el mejor modelo entre los dos o la predicción para la competencia no está bien hecha o la predicción en la competencia da menos de 0.5.

Detalles y recomendaciones

- Para consultas conceptuales sobre machine learning o preguntas de consigna pueden consultar en el canal de slack #consultas-tp3 o Piazza.
- Para consultas de código con su corrector o algún ayudante por privado.
- No deben buscar modelos entrenados por otros para usarlos, esto solo les puede jugar en contra porque es probable que no cumplan las condiciones pedidas, que no estén prolijos, que estén orientados a conseguir buenos resultados en la competencia (cosa que encima no evaluamos) y que tengan algún error conceptual.
- Recomendamos trabajar durante todo el TP en solo 4 notebooks: Uno de visualizaciones, otro para la regresión logística y uno para cada modelo de la parte III. Les recomendamos desarrollarlos de forma prolija y mostrar de forma ordenada cada uno de los resultados y pasos, con títulos y comentarios donde corresponda.
- El TP pide solo 6 visus y 3 modelos con condiciones muy claras, tengan esa consideración a medida avanzan para chequear que cumplen todo.
- El TP no pide ni evalúa más que lo que dice, si bien ser original y tener un buen score suma en términos de trabajo y aprendizaje para ustedes, sean inteligentes respecto a los modelos y features que eligen para trabajar para garantizar que pueden terminar. Ya van a tener tiempo de ser originales en el TP4…
- Particularmente este TP es muy difícil empezarlo al final, en cuotas se vuelve mucho más sencillo, les recomendamos empezar por las visus que no necesitan teoría nueva.
- Todos los puntos deben estar desarrollados (exceptuando por supuesto los extra).