PRUEBA
CIENTÍFICO DE DATOS

Prueba de habilidades y aptitudes Científico de datos Jr.

Dirección de Analítica
Circulode crédito

Tabla de Contenido

Introducción

El objetivo de esta prueba es demostrar tus habilidades desde los siguientes puntos de vista:

- Herramientas como R/PYTHON (a elección)
- Análisis de Datos
- Desarrollo de Modelos
- Argumentación
- Curiosidad e Investigación

Contesta este examen de manera individual. Puedes buscar información en sitios web si así lo requieres. Nos gustaría conocer cómo abordas los problemas, por favor documenta qué herramientas usaste.

A. Conocimientos de Riesgo de Crédito.

1. ¿Cuáles son las etapas en el ciclo del crédito?
2. Menciona algún indicador de riesgo de crédito.
3. Supongamos que un banco quiere desarrollar un modelo de score que le indique qué personas que solicitanun crédito van a caer en impago y cuáles no. Dicho banco desarrolló dos modelos, el ordenamiento de cadamodelo se muestra en las siguientes tablas:

Definiciones:

Campos tabla de odds	Definición
Rango Score	Los puntajes se dividen en intervalos del 10% de la población y se disponen en rangos de orden ascendente con los puntajes bajos en la parte superior y los altos en la inferior. Los puntajes altos indican un menor riesgo.
Personas	Muestra el número de expedientes (personas) en cada intervalo de puntajes.
Buenos	Muestra el número de expedientes (personas) que no cayeron en impago en cada intervalo de puntajes.
Malos	Muestra el número de expedientes (personas) que cayeron en impago en cada intervalo de puntajes.
Tasa de malos	Muestra el porcentaje de expedientes (personas) que cayeron en impago con respecto al total en cada intervalo de puntaje.
Malos acumulados	Muestra el porcentaje de expedientes (personas) que cayeron en impago con respecto al total en cada intervalo de puntajes mayores o iguales al del rango de puntaje considerado.

Modelo 1

	Personas	Buenos	Malos	Tasa de malos
Cartera	200,000	180,797	19,203	9.60%

Decil	Rango Score	Personas	Buenos	Malos	Tasa de Malos	Malos Acumulados
1	(300 , 445]	20000	15,673	4,327	21.64%	9.60%
2	(445 , 469]	20000	16,578	3,422	17.11%	7.44%
3	(469 , 492]	20000	17,034	2,966	14.83%	5.73%
4	(492 , 514]	20000	17,652	2,348	11.74%	4.24%
5	(514 , 539]	20000	17,999	2,001	10.01%	3.07%
6	(539 , 567]	20000	18,257	1,743	8.72%	2.07%
7	(567, 608]	20000	18,902	1,098	5.49%	1.20%
8	(608 , 649]	20000	19,302	698	3.49%	0.65%
9	(649 , 689]	20000	19,569	431	2.16%	0.30%
10	(689 , 818]	20000	19,834	169	0.85%	0.08%

Modelo 2:

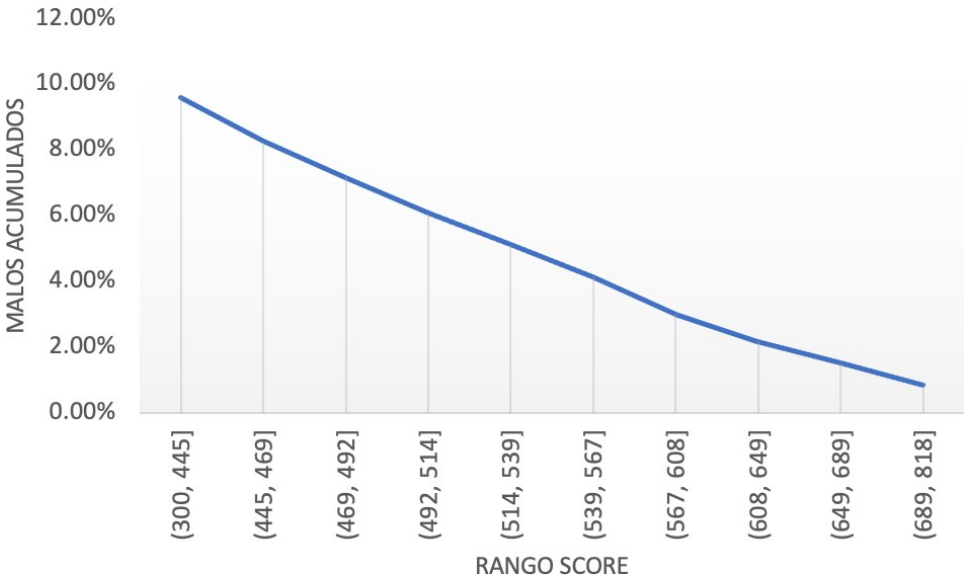
	Personas	Buenos	Malos	Tasa de malos
Cartera	200,000	180,797	19,203	9.60%

Decil	Rango Score	Personas	Buenos	Malos	Tasa de Malos	Malos Acumulados
1	(300 , 445]	20,000	15,673	4,327	21.64%	9.91%
2	(445 , 469]	20,000	17,178	2,822	14.11%	7.74%
3	(469 , 492]	20,000	16,634	3,366	16.83%	6.33%
4	(492 , 514]	20,000	15,998	4,002	20.01%	4.65%
5	(514 , 539]	20,000	17,999	2,001	10.01%	2.65%
6	(539 , 567]	20,000	18,600	1,400	7.00%	1.65%
7	(567, 608]	20,000	19,600	400	2.00%	0.95%
8	(608 , 649]	20,000	19,200	800	4.00%	0.75%
9	(649 , 689]	20,000	19,569	431	2.16%	0.35%
10	(689 , 818]	20,000	19,734	266	1.33%	0.13%

Justifica cuál de los dos modelos funciona correctamente.

4. Considerando la gráfica que se muestra a continuación, ¿qué estrategia de aceptación puede implementar el banco si su apetito de riesgo es que no más del 5% de su cartera caiga en impago?

Concepto	Definición
Malos acumulados	Muestra el porcentaje de expedientes (personas) que cayeron en impago con respecto al total en cada intervalo de puntajes mayores o iguales al del rango de puntaje considerado.



B. Conocimientos de Python y SQL

1. En un dataframe (df), ¿cuál es la sintaxis para obtener el total de na's para cada variable?

2. En ese mismo df, ¿cuál es la sintaxis para imputar los valores perdidos na con el promedio? Supón que los valores perdidos se encuentran en la variable 'edad' y 'nivel_socieco'.

3. Crea un histograma a través de la librería seaborn para la variable edad.

4. Estás a punto de preparar un modelo y para ello quieres separar en entrenamiento y validación. ¿Con qué sintaxis separarías tu base en un 70% para entrenamiento y el restante para validación? ¿qué librería usarías? Esto con una semilla aleatoria.

5. Si quisieras implementar un Random Forest 10 árboles y una máxima profundidad de 2. ¿Cuál sería la sintaxis? Puedes apoyarte de lo sig:

```
# Librería a importar
from sklearn import ….

# Inicializar el modelo
rf = …..(n_estimators = …..)
```

```
# Entrenar el modelo en tu base de entrenamiento: x_train, y_train
rf.fit(…. , ….)
```

6. Si quisieras verificar algunas métricas sobre tu modelo de Random Forest, ¿qué librería y sintaxis usarías para graficar la curva ROC? Supón que obtienes un valor de 0.53 de AUC, ¿este valor te diría que tu modelo es preciso?

7. Supón que tienes la sig información en las tablas uno y dos:

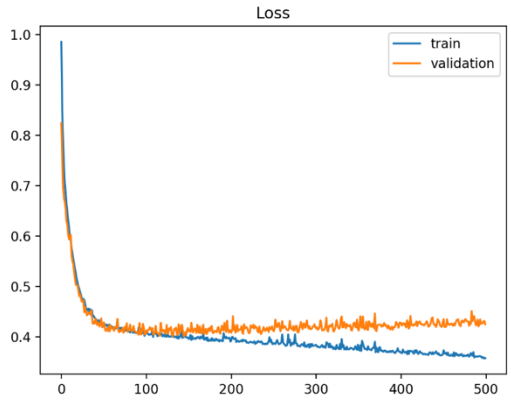
Tabla uno		
Año	Trim	Precio
2012	3	1200
2013	4	4000
2014	1	100

Tabla dos		
Año	Trim	Monto
2012	3	3.2
2013	4	5.6
2014	1	6.7

Escribe código sql para hacer un inner join de las tablas y seleccionar la suma de monto * precio por año.

C. Conocimientos de Ciencia de Datos

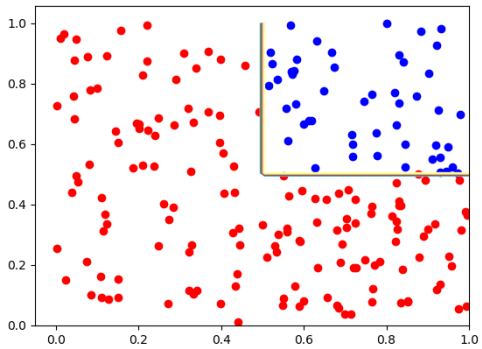
1. ¿Qué significan las siguientes curvas de aprendizaje? Justifica tu respuesta.



a) El modelo está sobre ajustado (overfit).
b) El modelo está sub ajustado (underfit).
c) El modelo está bien entrenado.
d) Ninguna de las anteriores.

2. Supón que tu modelo está sobre ajustado. ¿Qué harías para mejorar eso?

3. ¿Qué tipo de algoritmo realiza una separación como la que se muestra en la figura?



a) Máquina de Soporte Vectorial con Kernel Lineal
b) Regresión Logística
c) Árbol de Decisión
d) Random Forest

4. ¿Qué es el aprendizaje no supervisado? Menciona un ejemplo.

5. Si se te pidiera elaborar un modelo de machine learning de clasificación para un problema de fraude que tiene una bad rate del 2.0%, ¿con qué tipo de problema te podrías estar enfrentando? ¿cómo lo solucionarías? ¿qué tipo de modelo de ML usarías?

6. ¿Cuál consideras que es una buena métrica de performance para un problema desbalanceado, por decir, el 60% de AUCROC es una métrica buena o mala y porqué?

D. Problema en Python:

La empresa necesita realizar un modelo de fraude. Tu papel como DS es dar una solución de modelo y mitigar los riesgos de fraude basada en datos. Los datos para realizar el modelo son datos_fraud.csv .

1. Una vez que obtengas los datos, utiliza todos tus conocimientos y todos los pasos que creas necesarios para poder entregar un modelo funcional para predecir si un cliente debería de ser aprobado o no. Entre los pasos que esperamos ver están:
   - EDA
   - Pre-procesamiento de los datos
   - Entrenamiento del modelo
   - Testing de modelo
   - Una explicación de cómo pondrías este modelo en producción y que tendrías que estarle cuidando con el tiempo
2. Una vez entrenado tu modelo esperamos recibir 3 archivos específicos:
   - Un Jupyter Notebook que explique todo su proceso de entrenamiento. Aquí mismo es donde vas a incluir los distintos pasos descritos arriba bien documentados para poder entender cómo fuiste generando el modelo.
   - CSV de predicciones de tu modelo en testing data. El CSV nada más debe de incluir el ID de solicitante y su score del modelo.
   - Un PDF explicando qué punto de corte seleccionarías y por qué.

Recuerda que en un contexto de trabajo en equipo, las personas que leerán tu código puede que no estuvieron involucradas en su desarrollo pero igual tendrán que entendenderlo y/o mantenerlo.

Datos:
id: Identificador único de la transacción
timestamp: momento en el que ocurre la transacción
amount: monto de la transacción
variables_01 a variable_32: features de la trasacción
fraud: Flag de flaude 1 es transacción fraudulente y 0 no lo es