Logotipo, nombre de la empresa

Descripción generada automáticamente

**Prueba técnica**

Ciencia de Datos

Como parte de tu proceso se ha establecido esta prueba técnica donde podrás demostrar tus conocimientos y habilidades. Se plantea que logres resolver tareas de clasificación, regresión, o clustering con el tratamiento de variables apropiadas, los modelos que domines, así como la explicación e importancia en un contexto de negocio.

Considera el [conjunto de datos](https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients) (Default of Credit Card Clients) con los campos:

* **ID**: ID of each client
* **LIMIT\_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
* **SEX**: Gender (1=male, 2=female)
* **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
* **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
* **AGE**: Age in years
* **PAY\_0**: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
* **PAY\_2**: Repayment status in August, 2005 (scale same as above)
* **PAY\_3**: Repayment status in July, 2005 (scale same as above)
* **PAY\_4**: Repayment status in June, 2005 (scale same as above)
* **PAY\_5**: Repayment status in May, 2005 (scale same as above)
* **PAY\_6**: Repayment status in April, 2005 (scale same as above)
* **BILL\_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
* **BILL\_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
* **BILL\_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
* **BILL\_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
* **BILL\_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
* **BILL\_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
* **PAY\_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
* **PAY\_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
* **PAY\_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
* **PAY\_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
* **PAY\_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
* **PAY\_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
* **default.payment.next.month**: Default payment (1=yes, 0=no)

En donde debes llevar a cabo modelos para predecir la variable **PAY\_AMT4** y **default.payment.next.month** con los modelos que consideres apropiados.

Para esta prueba, se recomienda que lleves a cabo:

* Proceso de análisis exploratorio de los datos
* Realizar limpieza del dataset
* Creación de variables
* Modelado
  + Modelos lineales
  + Árboles de decisión
  + Modelos basados en boosting
  + Redes neuronales
* Evaluación
* Conclusiones a nivel técnico y a nivel de negocio

Una vez realizada tu solución, responde a las siguientes preguntas:

1. De las deficiencias en los datos, ¿cuáles y como las identificaste?
2. De realizar creación de variables, explica cuales hiciste y por qué.
3. De los modelos realizados, ¿cómo seleccionaste al mejor?
4. ¿Qué desafíos encontraste y cómo los superaste?

Entregables:

1. Presentación breve 2-3 slides con métricas como y conclusiones.

Consideraciones:

* Esta prueba debe hacerse en Python
* El tiempo para llevarse a cabo es de **72 horas** desde la recepción de la misma.
* Se sugieren modelos, no es necesario hacerlos todos.
* Siéntete libre de añadir pasos o análisis extra, estos serán un plus para tu proceso.
* El dataset es público, siéntete libre de subir tu código/solución a un repositorio
* Hint: El dataset contiene información de abril a septiembre, evita cometer “data leak” para con la variable continua a predecir.

Cuestionario general:

1. Establece con tus propias palabras, algunas buenas prácticas y funciones recomendadas para optimizar operaciones de lectura, escritura y manipulación en Spark/PySpark.
2. Indica las pruebas estadísticas que has utilizado como parte del desarrollo de una solución de ciencia de datos.
3. En el contexto de Machine Learning y Ciencia de datos, explica:
   1. No Free Lunch Theorem
   2. Occam’s Razor
   3. Data Leakage