



DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

Analítica Computacional para la toma de Decisiones – 202420

Proyecto 2 - Productos bancarios

**Integrantes: Jerónimo Vargas – 202113305, Carlos Gómez – 202111593,
Juan Esteban López – 202021417**

Asignación de Roles

Análisis de negocio – Carlos Gómez

Ingeniería de datos – Carlos Gómez

Análisis de datos – Jerónimo Vargas

Ciencia de datos – Jerónimo Vargas

Tablero de datos – Juan Esteban López

Despliegue – Juan Esteban López

Preguntas de negocio y plan de acción

Pregunta 1: ¿Qué segmentos de clientes muestran una mayor propensión a aceptar productos financieros?

Indicadores: Tasa de conversión por segmento: Este indicador mide la proporción de clientes en un segmento específico (por ejemplo, por ocupación, edad o balance) que aceptaron el producto financiero.

Visualizaciones: Gráfico de barras para la tasa de conversión por segmento: Muestra la tasa de conversión en cada segmento (ocupación, edad, etc.), con diferentes colores para alta, media y baja propensión.

Pregunta 2: ¿Es posible identificar a los clientes óptimos para aceptar productos financieros y así poder enfocar las campañas de ventas en aquellos con mayor probabilidad de aceptación, optimizando los recursos de marketing y aumentando las tasas de conversión?

Indicadores: Probabilidad de conversión del cliente: Este indicador se basa en la predicción del modelo (probabilidad de aceptación) y muestra la probabilidad de que cada cliente acepte un producto financiero.

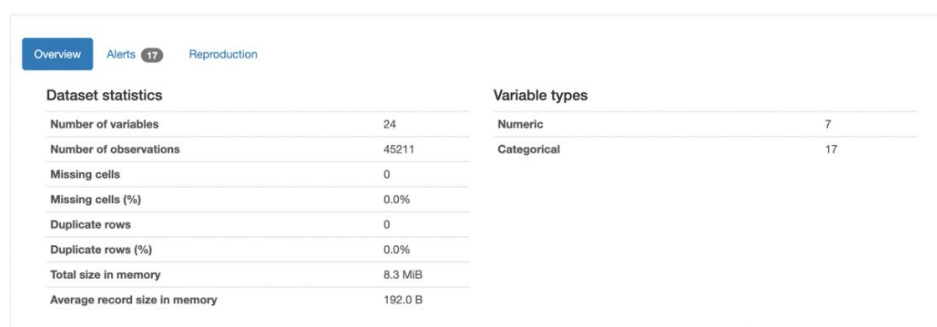
Score de propensión a la compra: Este indicador combina diversas características del cliente (balance, historial de contactos, ocupación, etc.) en un puntaje único que representa la probabilidad de aceptación de productos financieros.

Visualizaciones: Gráfico de distribución: Mostrar la distribución de las probabilidades de conversión de todos los clientes, agrupándolos en rangos de probabilidad. Gráfico de Dispersión: Un gráfico de dispersión donde el eje x represente el balance y el eje y el resultado de la campaña. Donde cada punto es un cliente.

Datos - Limpieza y alistamiento de datos

Todo el procedimiento realizado se puede encontrar en el documento de Jupyter Notebook “LimpiezaAlistamieto.ipynb”.

Primero, para realizar la exploración de los datos, se utilizó una librería llamada ydataprofiling (anteriormente conocida como pandas-profiling). Esta herramienta permite realizar un perfilamiento completo de los datos, proporcionando un análisis detallado que facilita la identificación de problemas como valores nulos, duplicados, tipos de datos incorrectos, y la distribución de las variables. A través de este perfilamiento, se puede evaluar la calidad del dataset y determinar qué pasos de limpieza y preprocesamiento son necesarios antes del análisis o modelamiento posterior. En la imagen proporcionada, se observa el resumen del perfil de los datos generados por ydata-profiling. Los principales hallazgos del reporte incluyen:



Overview	Alerts 17	Reproduction
Dataset statistics		
Number of variables	24	
Number of observations	45211	
Missing cells	0	
Missing cells (%)	0.0%	
Duplicate rows	0	
Duplicate rows (%)	0.0%	
Total size in memory	8.3 MiB	
Average record size in memory	192.0 B	
Variable types		
Numeric	7	
Categorical	17	

Imagen 1. Overview del perfilamiento de los datos.

Con este reporte, observamos que el conjunto de datos cumple con los principios de completitud y unicidad, ya que no tiene datos faltantes ni filas duplicadas. Además, no se han detectado alertas de posibles errores en el formato de los datos de cada variable. Por lo tanto, no fue necesario realizar procedimientos adicionales relacionados con estos aspectos.

A continuación, se procedió a realizar el alistamiento de datos:

Variable Job: En el análisis de la variable "Job" del dataset, se identificaron 12 valores distintos. Para simplificar y mejorar la interpretación de esta variable, se agruparon los trabajos en categorías generales según el tipo de labor que representan. Esta clasificación se realizó utilizando el sistema de "collar", que agrupa los trabajos en blue collar (trabajos manuales), white collar (trabajos administrativos), pink collar (trabajos en el sector de servicios y asistencia social), y otros. Posteriormente, se eliminó la variable original y se crearon nuevas variables binarias llamadas “Job_blue-collar”, “Job_white-collar” y “Job_pink-collar”. Estas variables indican con un valor de 1 si el trabajo de la persona pertenece a una de las categorías mencionadas. En los casos en que todas estas variables toman el valor de 0, se infiere que el trabajo de la persona corresponde a la categoría “other”. Este proceso permite representar de forma simplificada la clasificación de ocupaciones en el modelo.

Variable marital: Dado que la variable "Marital" es categórica y presenta tres valores posibles (Married, Single y Divorced), se optó por crear dos nuevas variables binarias llamadas “Marital_married” y “Marital_divorced”. En estos casos, un valor de 1 indica la categoría correspondiente. Cuando ambas variables toman el valor de 0, se infiere que la persona es Single.

Variable education: Dado que la variable "education" es categórica y presenta cuatro valores posibles (Primary, secondary, tertiary, unknown), se optó por crear tres nuevas variables binarias llamadas

“Education_primary”, “Education_secondary” y “Education_tertiary” . En estos casos, un valor de 1 indica la categoría correspondiente. Cuando estas variables toman el valor de 0, se infiere que la persona presenta un nivel de educación desconocido.

Variable default, housing, loan y y: Estas variables al ser categóricas con solo 2 valores posibles (Si y No) se optó por cambiarlas a una variable binaria donde el 1 representa el sí y el 0 representa el no.

Variable contact: Dado que la variable "Contact" es categórica y presenta tres valores posibles (telephone, cellulsr y unknown), se optó por crear dos nuevas variables binarias llamadas “Contact_cellular” y “Contact_unknown”. En estos casos, un valor de 1 indica la categoría correspondiente. Cuando ambas variables toman el valor de 0, se infiere que la persona fue contactada por celular.

Variable month: Dado que esta variable es categórica con 12 valores posibles de tipo cadena de texto (uno para cada mes del año), se transformaron estos valores en enteros del 1 al 12, donde el 1 representa enero y el 12 representa diciembre.

Variable poutcome: Dado que la variable "Poutcome" es categórica y presenta cuatro valores posibles (unknown, failure, sucess y other), se optó por crear tres nuevas variables binarias llamadas “Poutcome_failure”, “Poutcome_other” y “Poutcome_sucess”. En estos casos, un valor de 1 indica la categoría correspondiente. Cuando estas variables toman el valor de 0, se infiere que no se sabe la respuesta de la persona a campañas anteriores.

Variable pdays: Dado que la variable "Contact" contenía una gran cantidad de valores de 0 (indicando que los clientes no fueron contactados), se decidió eliminar la variable “Pdays”. Esto se debe a que esta variable no aportaba información adicional significativa y presentaba una alta autocorrelación con otra variable ya mencionada, lo cual reducía su utilidad para el modelo.

Exploración de datos

En esta sección, se realizará un análisis exploratorio de los datos para describir estadísticamente y visualizar el comportamiento de las diferentes variables presentes en el conjunto de datos. El objetivo es identificar patrones, tendencias y relaciones iniciales que permitan extraer primeras conclusiones relevantes sobre el conjunto de datos.

Estadísticas Descriptivas Clave:

Age: Rango amplio de edades, desde 18 hasta 95 años, con una media de 40.9 años. Esto indica una población diversa en términos de edad, que puede implicar diferentes necesidades y preferencias en productos financieros.

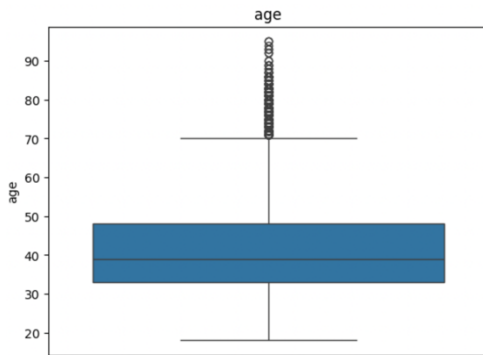
Balance: El balance promedio es de 1362.27, pero la desviación estándar es alta (3044.77), lo que sugiere una gran variabilidad en la situación financiera de los clientes. Además, el balance mínimo es negativo (-8019), lo cual indica que algunos clientes tienen deudas significativas.

Duration: La duración promedio de las llamadas es de 258 segundos, pero hay una variación considerable (desviación estándar de 257.53). Las llamadas más largas pueden indicar un mayor interés del cliente, con un máximo de hasta 4918 segundos.

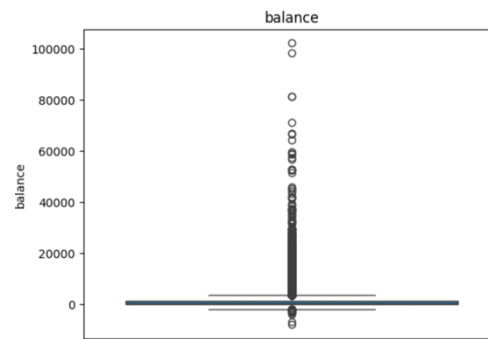
Marital: El 60.2% de los clientes están casados (marital_married), mientras que el 28.3% están solteros (marital_single). Este dato puede ser relevante para diseñar campañas dirigidas a distintos segmentos familiares.

Poutcome: El 10.8% de los clientes tuvo un resultado de "fracaso" (poutcome_failure) en campañas anteriores, mientras que el 4.1% tuvo un resultado positivo (poutcome_success). Esto sugiere que la mayoría de los clientes no respondió favorablemente en campañas previas, lo cual es un aspecto a considerar en estrategias futuras.

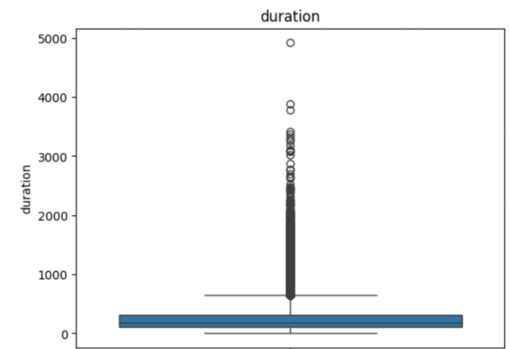
Diagramas de caja:



El diagrama de caja anterior muestra que la mayoría de los clientes tienen entre 30 y 50 años, lo cual representa el rango central. Hay varios valores atípicos por encima de los 60 años, que llegan hasta los 90. Estos outliers indican que aunque la mayoría de los clientes está en la edad laboral y en etapas de consolidación económica, también hay una minoría de clientes de edad avanzada, lo cual podría reflejar diferentes necesidades financieras y perfiles de riesgo.

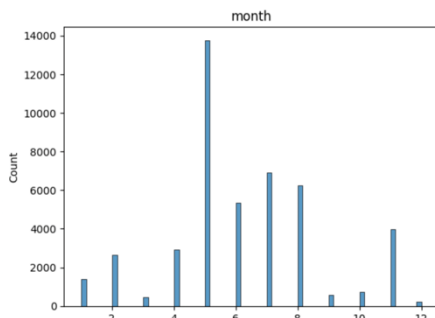


el diagrama de caja revela que la mediana es cercana a cero, indicando que una gran proporción de los clientes tiene balances bajos o incluso negativos. Sin embargo, hay una cantidad considerable de outliers en el extremo superior, con algunos balances que superan los 100,000. Esto sugiere la existencia de un pequeño grupo de clientes con balances significativamente altos, lo cual podría representar oportunidades para productos financieros específicos, como inversiones de alto valor o servicios premium.

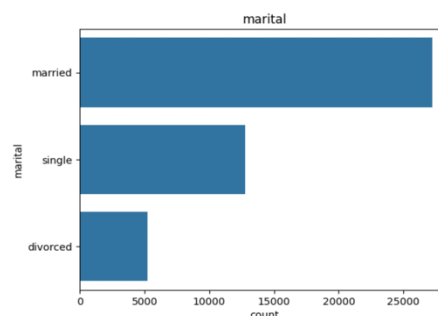


La duración de la llamada tiene una mediana baja, lo que indica que muchas de las interacciones con los clientes son cortas. No obstante, el diagrama muestra varios valores atípicos, con duraciones que llegan hasta los 4918 segundos. Estas llamadas excepcionalmente largas pueden ser una señal de un alto nivel de interés por parte de algunos clientes o de la complejidad de ciertos casos, lo cual podría ser relevante para entender el compromiso del cliente o para identificar prospectos de alta calidad.

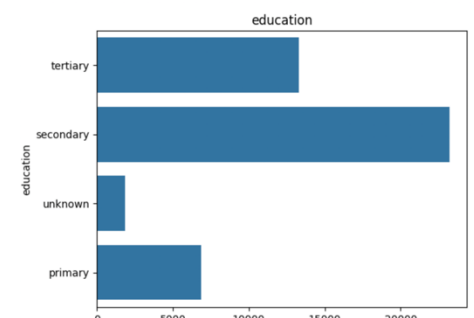
Histogramas:



En la variable mes de contacto, se observa que el número de contactos es significativamente mayor en algunos meses, con un pico notable en mayo y una menor frecuencia en meses como diciembre y marzo. Esto podría reflejar estrategias de campaña que se intensifican en ciertos períodos del año, posiblemente debido a tendencias de mercado o metas internas del banco. La concentración en algunos meses sugiere que existe una estacionalidad en las campañas de marketing, lo cual puede ser aprovechado para planificar y optimizar los esfuerzos de contacto, asegurando una distribución más equilibrada a lo largo del año y maximizando la efectividad en los períodos de alta actividad.



En la variable estado civil, se observa que la mayoría de los clientes están casados, seguidos por los solteros, mientras que los divorciados representan la menor proporción. Este patrón sugiere que una gran parte de la base de clientes puede estar en una etapa de vida establecida y con mayores responsabilidades financieras, lo cual podría hacerlos más receptivos a productos de inversión, seguros o planes de ahorro a largo plazo. En contraste, los clientes solteros podrían estar interesados en productos financieros con mayor flexibilidad o de corto plazo.



La distribución de la variable nivel educativo muestra que la mayoría de los clientes tiene educación secundaria, seguida de aquellos con educación terciaria. Un segmento menor cuenta con educación primaria, y un pequeño grupo tiene nivel educativo desconocido. Esta información es valiosa para personalizar las estrategias de comunicación y los productos financieros, ya que clientes con educación superior podrían estar interesados en productos de inversión más complejos o de mayor valor, mientras que aquellos con menor nivel educativo pueden preferir productos más simples y de bajo riesgo.

Diagramas de dispersión:

Balance vs. y: La mayoría de los clientes con altos balances no aceptaron el producto financiero ($y=0$), aunque se observa una pequeña cantidad de clientes con balances elevados que sí aceptaron ($y=1$). Esto sugiere que un alto balance no es un indicador definitivo de aceptación del producto.

Relación entre Default y y: La mayoría de los clientes que aceptaron el producto ($y=1$) no tenían un incumplimiento crediticio (default). Esto sugiere que la ausencia de incumplimientos es un factor positivo para la aceptación de productos financieros.

Relación entre Job (Blue Collar) y y: Los clientes con empleos de tipo "blue collar" muestran una mayor tendencia a no aceptar el producto financiero, aunque hay un grupo significativo que sí lo hace. Esto indica que este grupo puede necesitar estrategias específicas para mejorar la aceptación.

Modelos

Después de limpiar, preparar y analizar los datos, se buscó un método para estimar la respuesta de los clientes ante productos financieros. Para esto, se planteó utilizar un modelo de regresión lineal, ya que este tipo de modelo permite evaluar la relación entre variables independientes, como características demográficas y financieras de los clientes, y la variable dependiente, que en este caso representa la probabilidad de aceptación de un producto financiero. En este caso, se plantearon 3 diferentes modelos con el objetivo de determinar el mejor para estimar la variable independiente.

Modelo 1: Red más profunda con más capas ocultas y el optimizador SGD

Este modelo es una red neuronal secuencial implementada en Keras, diseñada para una tarea de clasificación binaria. La arquitectura del modelo incluye cuatro capas ocultas densamente conectadas, con tamaños de 128, 64, 32 y 16 neuronas respectivamente, cada una utilizando la función de activación ReLU para introducir no linealidad. La capa de salida tiene una sola neurona con activación sigmoide, adecuada para predecir probabilidades entre 0 y 1. El modelo utiliza el optimizador SGD y la función de pérdida de entropía cruzada binaria para el ajuste. Se entrena durante 100 épocas con un tamaño de lote de 32 y una validación del 20% de los datos de entrenamiento. En total, el modelo cuenta con 13,953 parámetros entrenables, lo cual permite ajustar adecuadamente el modelo para capturar patrones complejos en los datos.

Modelo 2: Red más ancha con Adam y función de activación tanh

Este modelo de red neuronal está diseñado con una arquitectura de tres capas ocultas, cada una con 256, 128 y 64 neuronas, respectivamente, y una función de activación tanh, que permite modelar relaciones no lineales en los datos. La capa de salida tiene una sola neurona con función de activación sigmoide, ideal para problemas de clasificación binaria. El modelo se entrena usando el optimizador Adam, lo que facilita una convergencia más rápida y precisa al ajustar los pesos. En total, el modelo tiene 142,085 parámetros, de los cuales 47,361 son entrenables, lo que indica una capacidad significativa para captar patrones complejos en los datos.

Modelo 3: Red más simple con el optimizador RMSprop

Este modelo tiene una estructura de red neuronal secuencial con dos capas ocultas de tamaños 32 y 16, cada una con la función de activación ReLU. La capa de salida cuenta con una neurona y utiliza la función de activación sigmoide, adecuada para problemas de clasificación binaria. El modelo se compila con el optimizador RMSprop, usando la función de pérdida de entropía cruzada binaria y la

métrica de precisión para evaluar su rendimiento. En total, el modelo cuenta con 2,628 parámetros, de los cuales 1,313 son entrenables y el resto corresponde a los parámetros del optimizador, lo que indica que es un modelo relativamente simple y eficiente.

Comparación de modelos:

Después de realizar los modelos se obtuvieron las medidas de desempeño para ver qué modelo era el que predecía de mejor forma la variable dependiente. En este caso se obtuvieron los siguientes resultados:

	Modelo	Accuracy	Precision	Recall	F1 Score
0	Modelo 1	0.891518	0.575368	0.294450	0.389546
1	Modelo 2	0.884994	0.512093	0.458137	0.483615
2	Modelo 3	0.894062	0.581649	0.351834	0.438453

Imagen 2. Tablas de indicadores modelos de redes neuronales.

Modelo 1 tiene una buena precisión (0.575), pero un recall bajo (0.294), lo cual indica que predice correctamente a algunos clientes que se suscribirán, pero pierde a muchos otros. Puede no ser la mejor opción si el objetivo es alcanzar a la mayor cantidad posible de clientes interesados.

Modelo 2 tiene el mejor balance entre precisión y recall, con un F1 Score de 0.483. Este modelo captura correctamente a más clientes que realmente se suscribirán (recall de 0.458), aunque también incluye algunos falsos positivos. Puede ser el mejor modelo si el área comercial quiere un equilibrio entre minimizar falsos positivos y no perder clientes potencialmente interesados.

Modelo 3 tiene la precisión más alta (0.581), lo que significa que sus predicciones positivas son más confiables, pero su recall es menor en comparación con el Modelo 2. Si la prioridad es evitar falsos positivos (es decir, predecir que un cliente se suscribirá solo cuando sea realmente probable), este modelo es adecuado. Sin embargo, sacrificará algo de cobertura, ya que no identificará a todos los clientes interesados.

Dado el contexto y la prioridad en la conversión de clientes, el Modelo 2 parece ser la mejor opción, ya que ofrece un equilibrio sólido entre precisión y recall, lo cual es útil para maximizar la conversión sin aumentar demasiado el número de falsos positivos.

Producto - Diseño y desarrollo del tablero

Para esta tarea primero se realizó el wireframe a mano para poder tener una idea general de cómo iba a quedar el tablero. Acá se tomó la decisión de pedir por parámetro información de los clientes; la información solicitada depende de la sección que se esté utilizando, esto se explica mejor más adelante. El wireframe se puede ver al final del documento en la sección de soportes.

En el producto final se le ofrece a la empresa una primera vista en la que se puede manipular lo que nosotros como grupo consideramos que son de las variables más importantes como el nivel de educación, la edad y la categoría en la que encaja el trabajo de la persona. En el contexto, creemos que estos son factores muy importantes para determinar si una persona puede aceptar o no un producto que le ofrezca nuestro cliente. Cuando el usuario de la aplicación cambia cambian dos gráficos; uno que es un gráfico que muestra la proporción de personas con esas características dadas que aceptaron productos financieros; el otro es un histograma que muestra el balance promedio que tienen estos clientes según

las características dadas. Adicional aparece un texto que da claridad de cuál fue el rango de edad escogido, la categoría del trabajo y el balance promedio de todas las personas que cumplen con la información dada por parámetro.

En la segunda vista que se le ofrece al usuario esta un modelo de clasificación de redes neuronales en el que el usuario ingresa todos los datos solicitados y el programa le retorna la probabilidad de que una persona con esas características acepte un producto financiero o no. Y con esa probabilidad el modelo también clasifica a la persona dependiendo de si alcanza un umbral establecido o no.

Tarea 6 - Despliegue

Para el despliegue primero se lanzó una base de datos relacional con el servicio de AWS denominado RDS. Para conectarnos a esta base de datos y poder alimentarla con un archivo CSV se lanzó una máquina virtual con el servicio EC2 de AWS, de esta forma, con las credenciales generadas previamente para la base de datos se pudo acceder y generar el esquema de tabla que se usó teniendo en cuenta la estructura del archivo CSV. El CSV que se tenía de forma local se pasó a la máquina virtual por consola para que después estando conectado desde la máquina virtual a la base de datos se pudiera copiar los datos del CSV en la tabla generada con su respectivo esquema previamente.

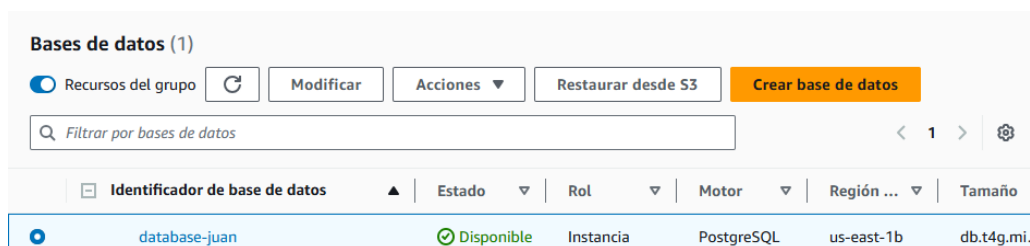


Imagen 3. Evidencia despliegue de la base de datos..

Luego para soportar el despliegue del tablero (DASH), se utilizó una máquina virtual (MV) en Amazon EC2 proporcionada a través de los recursos del laboratorio de aprendizaje de AWS Academy para desplegar el tablero de datos de la aplicación DASH. La instancia de EC2 configurada tiene las siguientes características: tipo de instancia t2.large, con una dirección IP pública elástica 98.82.44.247, ejecutando Amazon Linux.

El tablero DASH fue implementado en esta instancia para asegurar que estuviera accesible y ejecutándose de manera continua para los usuarios interesados en visualizar los resultados de análisis de datos de manera interactiva y efectiva. Todo el repositorio, incluyendo los archivos necesarios para soportar la funcionalidad del tablero, fue cargado desde GitHub a la instancia de EC2. Esto incluye el modelo serializado en formato .h5 y los datos de entrenamiento en formato .npy, que son utilizados para escalar los parámetros de entrada en el tablero DASH, garantizando que los datos introducidos por el usuario se procesen de manera coherente con el modelo entrenado.

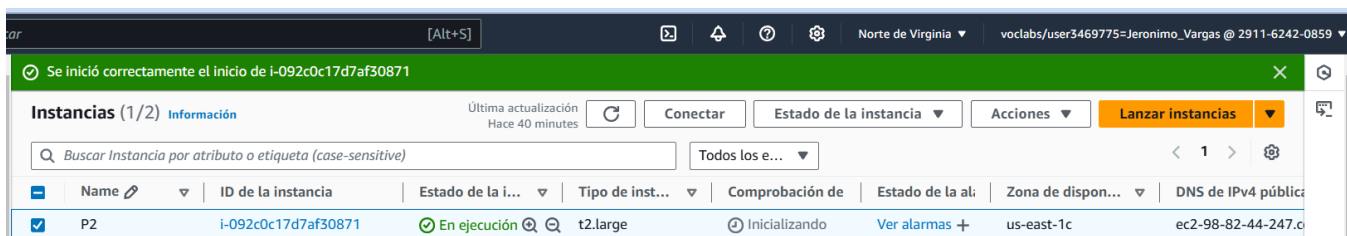
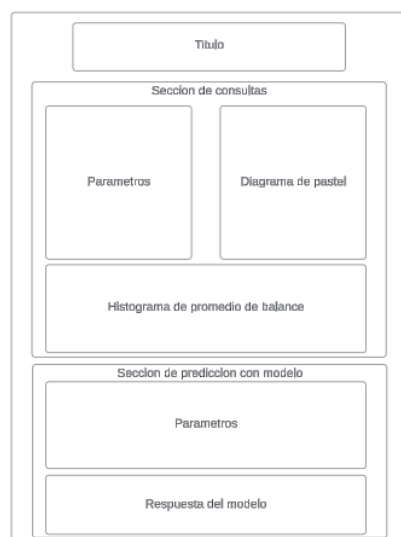


Imagen 1. Overview del perfilamiento de los datos.

Anexos

Wireframe DASH:



Reporte trabajo en equipo

En este proyecto, trabajamos de manera colaborativa para asegurarnos de que todos los miembros del equipo comprendieran las tareas que se estaban realizando en cada etapa del proceso. Aunque todos contribuimos al entendimiento general del proyecto, cada miembro asumió un rol de liderazgo en las áreas específicas seguridad de acuerdo con sus habilidades y conocimientos. Esta dinámica nos permitió ser más eficientes y asegurarnos de que cada tarea fuera manejada por la persona más adecuada.

Jerónimo Vargas Rendon: En el Proyecto 2, me encargué del análisis y la ciencia de datos. Realicé la exploración inicial del dataset usando herramientas de perfilamiento y visualizaciones para identificar patrones clave y preparar los datos. Desarrollé y evalué tres modelos de redes neuronales, eligiendo el Modelo 2 por su balance entre precisión y recall. Además, configuré la preparación de datos para la predicción en tiempo real en el tablero DASH, asegurando su funcionalidad.

Carlos Gómez: En el proyecto, mi principal contribución se centró en la limpieza y preparación de los datos, analizando cuidadosamente cada variable para transformar la información de manera que se minimicen sesgos y se conserve toda la información relevante para el modelo. Además, fui responsable de formular preguntas de negocio y desarrollar indicadores y visualizaciones que proporcionan una mejor comprensión de los datos y apoyan la toma de decisiones estratégicas.

Juan Esteban López: Mi aporte en el proyecto fue la realización del tablero en el que se muestran los datos al usuario. También realice el despliegue de la aplicación en la máquina virtual en Amazon Linux (servicio EC2) y la base de datos en Postgresql (servicio RDS). El despliegue del DASH con su respectiva ip elástica.

Soportes

- Soporte 1: análisis de negocio (soporte incluido en el reporte).
- Soporte 2: fuentes de limpieza “LimpiezaAlistamiento.ipynb”
- Soporte 3: fuentes de análisis “Exploracion.ipynb”
- Soporte 4: fuentes de modelización “ModeloRedesN.ipynb”
- Soporte 5: fuentes del tablero Dash.py

- Enlace despligue: <http://98.82.44.247:8050>

Estudio de aceptación de productos bancarios

Seleccione el trabajo que desea estudiar

Tenga en cuenta que los trabajos están divididos en 3 categorías:

- Blue Collar: Trabajos manuales o de fábrica
- Pink Collar: Trabajos de servicios
- White Collar: Trabajos de oficina o gerenciales

Blue Collar

Seleccione el rango de edad

0 10 20 30 40 50 60 70 80 90 100

Seleccione el nivel de educación

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

Tasa de Conversión de Clientes

Categoría	Tasa de Conversión
0 (Green)	94.3%
1 (Orange)	5.75%

- Soporte 7: repositorio Git en Github <https://github.com/Jeronimo2122/P2.git>