Resumen del Análisis del Notebook

El notebook analiza un dataset relacionado con una campaña de marketing portuguesa enfocada en la venta de préstamos bancarios. Desde el inicio, se plantea el objetivo de limpiar y explorar los datos para maximizar el valor de la información obtenida. La estructura del notebook está organizada de manera lógica, comenzando con la carga y revisión inicial del dataset utilizando PySpark. La lectura del archivo CSV incluye la inferencia de esquemas, el uso de encabezados y el análisis preliminar de dimensiones mediante el conteo de filas y columnas.

En la fase inicial, se verifica la distribución de la variable objetivo (target), destacando un desbalance en los datos: solo el 11.7% de los clientes contrataron depósitos a plazo fijo. Esto sugiere la necesidad de tratar el desbalance si se planea utilizar modelos predictivos. Posteriormente, el análisis se centra en la calidad de los datos mediante la identificación de variables categóricas y numéricas. Aquí se realiza una inspección visual de las primeras filas, se verifica el esquema del dataset y se calculan estadísticas descriptivas básicas.

El proceso de limpieza incluye la detección y eliminación de duplicados. Tras un análisis exhaustivo, se concluye que no hay registros duplicados significativos que requieran eliminación adicional. También se examinan las columnas categóricas para identificar aquellas que no aportan información útil. Se resalta el caso de la columna "default", que presenta un desbalance extremo y podría no ser relevante para análisis futuros. Este tipo de evaluación asegura que las variables utilizadas sean significativas y reduzcan el ruido en los resultados.

Por último, se realiza una exploración profunda de las variables categóricas, analizando sus valores únicos y distribuciones. Este enfoque permite identificar patrones o anomalías que podrían ser relevantes para las siguientes etapas del análisis. En general, el notebook refleja un enfoque ordenado y detallado para la exploración y preparación de datos, sentando las bases para un análisis más avanzado o modelado predictivo. Si se complementa con técnicas de tratamiento de desbalance y análisis estadístico, podría proporcionar resultados robustos y aplicables en un contexto real.