Informe de Análisis Exploratorio de Datos (EDA) – Parcial 2

1. Objetivo del Análisis

El propósito de este EDA es comprender la estructura, calidad y comportamiento de las variables incluidas en las bases de datos entregadas para el Parcial 2 de Finanzas Descentralizadas IV, con el fin de establecer una base sólida para el desarrollo del **Modelo de Machine Learning (Punto 1 y 2)**.

Las bases utilizadas fueron:

- `base_modelo_40k.csv`: 40.000 registros con variable objetivo `CLIENTE_MORA`.
- `base_prueba_10k_sin_mora.csv`: 10.000 registros sin variable objetivo (para predicción y validación final).

---

2. Exploración Inicial y Estructura

- **Tamaños**: 40.000 filas x 15 columnas (train) y 10.000 filas x 14 columnas (test).
- **Diferencia de esquema**: la única columna ausente en test es `CLIENTE_MORA`, correspondiente a la variable dependiente.
- **Variables principales**: `EDAD`, `INGRESO`, `SCORE_DATACREDITO`, `SALDO_ROT`, `SALDO_FIJO`, `SALDO_SF`, `CRED_REESTRUCTURADO`, `TIENE_HIPOTECA`, `ESTADO_MORA_REAL`, `ESTADO_MORA_FIN`, entre otras.
- **Tipos de datos**: predominan las numéricas (continuas o proporciones), y unas pocas categóricas binarias o discretas (`SEXO`, `CRED_REESTRUCTURADO`, `TIENE_HIPOTECA`).
- **Valores nulos**: proporciones insignificantes en todas las variables (<1%), no requieren imputación especial.

---

3. Variable Objetivo: `CLIENTE_MORA`

- Distribución: 36.855 clientes sin mora (0) y 3.145 con mora (1).
- **Tasa de mora:** 7,86% del total.
- **Implicación**: dataset **altamente desbalanceado**, por lo que el modelo requerirá técnicas de **reponderación o balanceo** (por ejemplo, `scale_pos_weight` o SMOTE).

---

4. Análisis Descriptivo de Variables Numéricas

- `INGRESO`: distribución sesgada a la derecha (cola larga). Los ingresos bajos están sobrerrepresentados.

- `EDAD`: concentrada entre 30 y 50 años.
- `SCORE_DATACREDITO`: se concentra entre 700 y 900 puntos.
- `SALDO_ROT`, `SALDO_FIJO`, `SALDO_SF`: presentan colas largas; algunos outliers en valores altos.
- **Conclusión general:** la población tiene perfiles de ingresos y scores variados, lo que podría ayudar a segmentar bien el riesgo crediticio.

---

5. Correlación con la Variable Objetivo (`CLIENTE_MORA`)

Se calcularon correlaciones punto-biseriales (equivalentes a Pearson con objetivo binario). Las variables con mayor asociación fueron:

Variable	Correlación (aprox.)	Interpretación
`SCORE_DATACREDITO`	-0.27	Score alto → menor probabilidad de mora.
`ESTADO_MORA_FIN`	+0.13	Mora final alta → mayor riesgo.
`SALDO_ROT`	+0.11	Saldos rotativos altos correlacionan con mora.
`CRED_REESTRUCTURADO`	+0.09	Clientes reestructurados tienden a mayor mora.
`INGRESO`	-0.04	Ingreso mayor reduce leve riesgo.
`EDAD`	-0.03	Clientes jóvenes presentan mayor mora relativa.

Conclusión: el score crediticio y las variables de comportamiento financiero son los principales predictores.

---

6. Variables Categóricas

- `SEXO`: proporción equilibrada (hombres/mujeres).
- `CRED_REESTRUCTURADO`: minoría significativa con impacto positivo en mora.
- `TIENE_HIPOTECA`: la mayoría no posee hipoteca.
- **Relaciones mixtas (boxplots)**: ingresos y scores difieren ligeramente entre clases, pero sin separación perfecta.

---

7. Comparación Train (40k) vs Test (10k)

- **Distribuciones** similares en medias y desviaciones estándar.
- No se observan desviaciones relevantes (**no hay data drift**), por lo que el modelo puede generalizar adecuadamente.

---

8. Hallazgos Principales

1. Dataset limpio, sin nulos significativos.
2. Fuerte desbalance en la variable objetivo (7,86% mora).
3. Variables predictivas relevantes: `SCORE_DATACREDITO`, `SALDO_ROT`, `ESTADO_MORA_FIN`, `CRED_REESTRUCTURADO`.
4. Las variables de ingresos y edad influyen, pero con menor peso.
5. No hay sesgos evidentes entre el conjunto de entrenamiento y testeo.

---

9. Conclusiones y Recomendaciones para el Punto 2

- Se confirma la viabilidad del dataset para modelado supervisado.
- El próximo paso debe abordar el **balanceo de clases** y la **optimización de hiperparámetros** en un modelo robusto como **XGBoost**.
- Se sugiere evaluar **umbral de probabilidad** para mantener una tasa de mora proyectada $\leq$ 2.5%, de acuerdo con la política de riesgo establecida.
- Las métricas clave a comparar serán: **Recall**, **Precision**, **F1**, **ROC-AUC**, y la diferencia entre **Recall_test** y **Recall_10k** (control de sobreajuste).

Estado final: el EDA evidencia coherencia, limpieza y potencial predictivo adecuado; se recomienda proceder con el desarrollo del **Modelo de Machine Learning (Punto 2)** bajo criterios de interpretabilidad y rentabilidad financiera.