# Tinvest - Análisis Exploratorio de Datos (EDA)

## 📋 Descripción General

Este repositorio contiene un análisis exploratorio de datos (EDA) completo para una plataforma fintech de inversión. El análisis se enfoca en entender el comportamiento de los clientes, identificar patrones de crecimiento (NNM - Net New Money), analizar el riesgo de churn y desarrollar estrategias de retención basadas en datos.

**Fecha de corte del análisis:** 31 de diciembre de 2024
**Base de clientes:** 800 clientes únicos
**Período analizado:** Transacciones y balances desde 2021 hasta 2024

---

## 📁 Estructura del Repositorio

```
tinvest/
├── data/
│   ├── clients.csv              # Información demográfica de clientes
│   ├── transactions.csv         # Histórico de transacciones
(depósitos/retiros)
│   └── portfolio_balance.csv    # Saldos mensuales por cliente y producto
├── notebooks/
│   ├── eda.ipynb                # Notebook principal de análisis
│   └── plots_fintech_html/      # Gráficos interactivos generados (HTML)
├── utils/
│   └── utils.py                 # Funciones utilitarias para
visualización y análisis
└── docs/
    ├── tin_investing.pdf
    └── tin_investing.pptx
```

---

## 🔍 Análisis Realizado en el Notebook EDA

### 1. Análisis de NNM (Net New Money) y AUM (Assets Under Management)

**Objetivo:** Entender la salud financiera de tin tinvest y los patrones de crecimiento.

**Métricas calculadas:**

- **NNM mensual:** Diferencia entre depósitos y retiros por mes
- **AUM mensual:** Saldo total gestionado por mes
- **Churn:** Churn que representa la permanencia y lealtad de los inversionistas.

**Hallazgos principales:**

- ✅ Todos los meses presentan NNM positivo (no hay fuga neta de fondos)

- 📈 El AUM crece de forma monótona sin caídas relevantes
- 🎯 Se identificaron 2 meses pico con NNM excepcional:
  - **Septiembre 2024:** 1,725 millones COP (89.6% en ACCIONES, 96% segmento premium)
  - **Marzo 2024:** 1,497 millones COP (89% en FIC, 68.7% segmento premium)

**Conclusión:** La plataforma está en fase de crecimiento saludable. Los picos de NNM representan oportunidades para documentar buenas prácticas comerciales y replicarlas sistemáticamente.

---

## 2. Descomposición de Meses Pico de NNM

**Análisis realizado:**

- Desglose por producto (ACCIONES, FIC, CDT, FPV)
- Desglose por segmento (premium vs retail)
- Análisis de nuevos vs clientes existentes
- Análisis de Pareto (concentración de NNM por cliente)

**Hallazgos clave:**

**Septiembre 2024:**

- 15.5% de los clientes explica el 80% del NNM
- 100% del NNM proviene de clientes existentes (no hubo nuevos clientes significativos)
- ACCIONES dominó con 89.6% del NNM

**Marzo 2024:**

- 18.1% de los clientes explica el 80% del NNM
- 5.1% del NNM provino de clientes nuevos
- FIC dominó con 89% del NNM

---

## 3. Análisis de Concentración y Estrategia de NNM

**Métricas calculadas:**

- **Pareto de clientes:** Identificación del % de clientes que aporta el 80% del NNM
- **Intensidad de inversión:** Ratio NNM / Ingreso mensual
- **Retención de capital por segmento:** (Depósitos - Retiros) / Depósitos
- **Productos estrella:** Ranking de productos por NNM neto

**Hallazgos principales:**

- 📊 **39% de los clientes** (312 clientes) aporta el **80% del NNM total**
- 💰 **Productos por NNM neto:**
  1. FIC: 4,875 millones COP
  2. CDT: 2,773 millones COP
  3. ACCIONES: 1,886 millones COP
  4. FPV: 817 millones COP
- 🎯 **14 clientes ricos** están sub-invertidos (ingresos > percentil 75, pero intensidad < 10%)

- 📉 **Retención de capital:**
  - Retail: 63.0%
  - Premium: 48.7% (mayor rotación de capital)

**Oportunidad identificada:** "Ballenas sin invertir" - clientes con alto poder adquisitivo pero baja intensidad de inversión.

---

## 4. Análisis de Vintages (Año de Inicio)

**Objetivo:** Entender la calidad y antigüedad del saldo por producto.

**Métricas calculadas:**

- Antigüedad promedio de la relación cliente-producto (en días)
- Composición del saldo por año de inicio (vintage)
- Historia del producto (fechas de primera y última transacción)

**Insights:**

- Permite identificar qué productos tienen relaciones más longevas
- Muestra la distribución temporal de los saldos actuales
- Útil para estrategias de retención basadas en antigüedad

---

## 5. Análisis de Churn Operativo

**Enfoque data-driven:** En lugar de usar umbrales arbitrarios (ej: 90 días), se definen umbrales basados en los propios datos.

### 5.1 Curva de Retorno (Latencia entre Transacciones)

**Método:**

- Analiza los gaps (pausas) entre transacciones consecutivas de cada cliente
- Calcula la probabilidad acumulada de retorno según días de silencio
- Identifica el umbral donde el 90% de los clientes ya ha retornado

**Resultados:**

- Tiempo promedio de silencio: **20 días**
- Se analizaron **21,420 pausas** entre transacciones
- Umbral sugerido basado en probabilidad de retorno del 90%

### 5.2 Sensibilidad de Saldo

**Método:**

- Analiza la tasa de inactividad por rango de saldo
- Identifica el umbral de saldo donde ≥90% de los clientes están inactivos

**Resultados:**

- Permite definir un umbral de saldo "efectivamente cero" para churn
- Combina días de inactividad + saldo bajo para una definición más precisa

**Definición final de churn operativo:**

- Días sin transaccionar ≥ umbral calculado (basado en curva de retorno)
- Saldo actual ≤ umbral calculado (basado en sensibilidad)

**Tasa de churn operativo:** 13.8% (110 clientes de 800)

---

## 6. Perfilamiento de Clientes con Churn

**Análisis comparativo:** Características demográficas y de comportamiento de clientes con churn vs retenidos.

**Dimensiones analizadas:**

1. **Poder adquisitivo:** Distribución de ingresos mensuales
2. **Edad:** Distribución de edades (KDE)
3. **Apetito de riesgo:** Distribución de risk_score (violin plot)
4. **Churn por segmento:** Tasa de churn por segmento de negocio

**Hallazgos:**

- Permite identificar qué perfiles de clientes son más propensos al churn
- Facilita el diseño de estrategias de retención segmentadas

---

## 7. Modelo Predictivo de Churn con Estrategia de Retención

**Objetivo:** Predecir qué clientes valiosos están en riesgo de churn para priorizar esfuerzos de retención.

### 7.1 Construcción del Modelo

**Features utilizadas:**

- `income_monthly`: Ingreso mensual
- `risk_score`: Score de riesgo del cliente
- `avg_aum`: Saldo promedio administrado
- `days_since_txn`: Días desde última transacción
- `txn_count`: Número total de transacciones
- `tenure_months`: Antigüedad en meses

**Target:** `is_churn` (saldo actual < 50,000 COP)

**Modelo:** Random Forest Classifier

- **ROC-AUC Score:** 0.756
- Clase balanceada para manejar desbalance

### 7.2 Matriz de Estrategia: Riesgo vs Valor

**Segmentación en 4 cuadrantes:**

1. **PRIORIDAD CRÍTICA (Salvar):** Alto riesgo de churn + Alto valor
    - 40 clientes identificados
    - Acción: Campañas de retención urgentes
2. **FIDELIZAR (Cuidar):** Bajo riesgo + Alto valor
    - 760 clientes identificados
    - Acción: Programas de fidelización y cross-sell
3. **DEJAR IR (No rentable):** Alto riesgo + Bajo valor
    - 480 clientes identificados
    - Acción: No invertir recursos en retención
4. **BAJA PRIORIDAD:** Bajo riesgo + Bajo valor
    - 1,920 clientes identificados
    - Acción: Mantener con esfuerzos mínimos

**Valor estimado:** Proxy de CLV (Customer Lifetime Value) = AUM promedio × 1.5%

---

## 🛠️ Funciones Principales en `utils.py`

### Visualización

- `plot_nnm_vs_aum()`: Gráfico de NNM mensual vs evolución de AUM
- `plot_nnm_peak_summary()`: Dashboard 2x2 con descomposición de meses pico
- `plot_return_curve()`: Curva de latencia de retorno de clientes
- `plot_balance_sensitivity()`: Sensibilidad de inactividad por rango de saldo
- `plot_churn_profile_subplots()`: Perfilamiento de clientes churn vs retenidos
- `plot_product_vintage_dashboard_plotly()`: Dashboard de vintages por producto
- `plot_nnm_strategy_dashboard()`: Dashboard estratégico de NNM
- `plot_feature_importance_plotly()`: Importancia de variables del modelo
- `plot_risk_value_matrix_plotly()`: Matriz de estrategia riesgo vs valor

### Utilidades

- `apply_corporate_layout()`: Aplica estilo corporativo a gráficos Plotly
- `save_html()`: Guarda gráficos como HTML interactivos

---

## 📊 Conclusiones y Hallazgos Principales

### 1. Salud Financiera Sólida

- ✅ Crecimiento sostenido: NNM positivo todos los meses
- ✅ AUM en crecimiento constante sin caídas significativas
- ✅ No hay "hemorragia" de fondos

### 2. Concentración de Valor

- 📊 Principio de Pareto aplicado: 39% de clientes aporta 80% del NNM
- 🎯 Oportunidad de crecimiento: 14 clientes ricos sub-invertidos identificados
- 💰 Segmento premium domina en meses pico (96% en sep-2024)

### 3. Productos Líderes

- 🥇 **FIC** es el producto estrella por NNM neto (4,875M COP)
- 🥈 **CDT** en segundo lugar (2,773M COP)
- 📈 **ACCIONES** tuvo un mes excepcional en sep-2024

### 4. Retención de Capital

- ⚠️ Segmento premium tiene menor retención (48.7%) vs retail (63.0%)
- 💡 Oportunidad: Mejorar retención en segmento premium

### 5. Churn Operativo

- 📉 Tasa de churn: 13.8% (110 clientes)
- 🔬 Definición basada en datos (no arbitraria)
- ⏱️ Tiempo promedio de silencio: 20 días

### 6. Modelo de Retención Estratégica

- 🎯 40 clientes en "PRIORIDAD CRÍTICA" requieren acción inmediata
- 💎 760 clientes valiosos para fidelizar
- 📊 Modelo con ROC-AUC de 0.756 (buen desempeño)

### 7. Oportunidades de Crecimiento

- 🐋 "Ballenas sin invertir": 14 clientes con alto poder adquisitivo pero baja inversión
- 📈 Replicar estrategias de meses pico (mar-2024 y sep-2024)
- 🎯 Enfoque en clientes existentes (no solo nuevos)

---

## 🚀 Cómo Usar Este Repositorio

### Requisitos

```
pip install pandas numpy matplotlib seaborn plotly scikit-learn shap
```

### Ejecutar el Notebook

1. Abrir `notebooks/eda.ipynb` en Jupyter Notebook o JupyterLab
2. Asegurarse de que los datos estén en `data/`
3. Ejecutar las celdas en orden
4. Los gráficos se guardarán automáticamente en `notebooks/plots_fintech_html/`

### Estructura de Datos Esperada

**clients.csv:**

- `client_id`: ID único del cliente
- `registration_date`: Fecha de registro
- `age`: Edad
- `income_monthly`: Ingreso mensual
- `segment`: Segmento (premium/retail)
- `risk_score`: Score de riesgo (0-1)

**transactions.csv:**

- `client_id`: ID del cliente
- `date`: Fecha de transacción
- `product`: Producto (ACCIONES, FIC, CDT, FPV)
- `type`: Tipo (deposit/withdrawal)
- `amount`: Monto

**portfolio_balance.csv:**

- `client_id`: ID del cliente
- `date`: Fecha del balance
- `product`: Producto
- `balance`: Saldo

---

## 📈 Próximos Pasos Recomendados

1. **Documentar playbooks comerciales** de los meses pico (mar-2024 y sep-2024)
2. **Campaña de retención** para los 40 clientes en "PRIORIDAD CRÍTICA"
3. **Programa de fidelización** para los 760 clientes valiosos
4. **Estrategia de cross-sell** para los 14 "ballenas sin invertir"
5. **Mejorar retención de capital** en segmento premium
6. **Monitoreo continuo** del modelo de churn (reentrenar periódicamente)

---

## 📝 Notas Técnicas

- **Lenguaje:** Python 3.x
- **Librerías principales:** pandas, numpy, plotly, scikit-learn
- **Visualizaciones:** Plotly (interactivas, guardadas como HTML)
- **Modelo:** Random Forest con class_weight="balanced"
- **Métricas:** ROC-AUC para evaluación del modelo

---

## 👥 Autor

Juan David Rincón

---