Informe de Análisis Exploratorio de Datos

Caja de Ahorros - Proyecto de Predicción de Ingresos

Versión del documento: 1.0
Fecha: Septiembre 2025
Dirigido a: Alta Dirección, Equipo de Ciencia de Datos y Stakeholders de Negocio

Resumen Ejecutivo

Este informe presenta el análisis exploratorio de datos (EDA) realizado sobre la base de clientes para el modelo de predicción de ingresos. El análisis de **29,319 clientes únicos** reveló hallazgos claves que guiaron la estrategia de modelado y los estándares de calidad de datos.

Hallazgos Principales

- **Calidad de Datos:** Se consolidaron 42,549 registros en 29,319 clientes únicos
- **Optimización de Variables:** Reducción de la complejidad categórica en un 98.5%, manteniendo la relevancia de negocio
- **Cobertura Alcanzada:** Cobertura de datos entre 60-80% tras la simplificación de variables categóricas
- **Preparación para Producción:** Establecimiento de estándares robustos de calidad de datos para uso operacional

Descripción General del Dataset

Fuentes de Datos y Consolidación

Fuente	Registros	Clientes Únicos	Cobertura
Info_Cliente.csv	19,047	19,047	Dataset principal
Info_Clientes_2.csv	23,502	23,502	Dataset secundario
Consolidado Final	**42,549**	**29,319**	**100%**

Categorías de Variables

Se identificaron **24 variables clave** agrupadas en cuatro categorías principales:

Demográficos (6 variables)

- ID de cliente e identificador único
- Edad, género, estado civil
- Localización geográfica (ciudad, país)

Información Laboral (4 variables)

- Ocupación y cargo
- Nombre del empleador y fecha de inicio laboral

Perfil Financiero (8 variables)

- Saldo de cuenta y pagos mensuales
- Monto de préstamos y tasas de interés
- Uso de productos e historial de pagos

Variables Temporales (6 variables)

- Fechas de inicio y cierre de cuentas
- Antigüedad laboral
- Fechas y horas de procesamiento de datos

Desafíos Críticos de Calidad de Datos

Desafío 1: Alta Cardinalidad Categórica

El análisis inicial reveló una cardinalidad extremadamente alta en varias variables categóricas:

Variable	Categorías Originales	Impacto en el Negocio
Nombres de Empleadores	7,698 valores únicos	83% aparecen una sola vez
Cargos Laborales	2,178 valores únicos	72% aparecen una sola vez
Ocupaciones	245 valores únicos	Complejo pero manejable
Ciudades	78 valores únicos	Diversidad geográfica

Implicación de Negocio: Sin tratamiento adecuado, se generarían más de 10,000 variables en el modelo, provocando:

- Predicciones poco confiables por baja muestra por categoría
- Ineficiencia de memoria y cómputo
- Dificultad en interpretación y mantenimiento del modelo

Desafío 2: Inconsistencias en Registro de Datos

Se identificaron problemáticas frecuentes de calidad:

- **Variaciones de mayúsculas/minúsculas:** “JUBILADO”, “jubilado”, “Jubilado”
- **Caracteres especiales:** “POLICÍA” vs “POLICIA”
- **Sinónimos:** “PROFESOR” vs “DOCENTE” vs “MAESTRO”

- **Espaciado:** Espacios adicionales e inconsistencias de formato

---

Solución Estratégica: Consolidación Inteligente de Categóricos

Enfoque: Estrategia “Top-N + Otros”

En lugar de codificación tradicional (one-hot, label) que crearía miles de variables, se implementó una consolidación orientada al negocio:

1. **Identificar las categorías top** con mayor valor para el negocio
2. **Agrupar el resto** en “Otros” estandarizados
3. **Mantener 60-80% de cobertura** tras simplificar variables
4. **Reglas de codificación seguras para producción**

Resultados: 98.5% de Reducción de Complejidad

Variable	Antes	Después	Reducción	Cobertura
Empleadores	7,698	7 categorías	99.9%	60%
Cargos	2,178	7 categorías	99.7%	60%
Ocupaciones	245	7 categorías	97.1%	39%
Ciudades	78	6 categorías	92.3%	80%
Total	**10,199**	**29 categorías**	**98.5%**	**60-80%**

Categorías de Negocio Aprobadas

Categorías de Ocupación

Ocupaciones Principales (Top 6):

- JUBILADO - 16.6% de clientes
- DOCENTE - 7.1% de clientes
- POLICIA - 5.4% de clientes
- OFICINISTAS - 3.7% de clientes
- SUPERVISOR - 3.6% de clientes
- ASISTENTE - 3.0% de clientes

Principales Empleadores (Top 6):

- NO APLICA - 15.1%
- MINISTERIO DE EDUCACION - 8.3%

- MINISTERIO DE SEGURIDAD PUBLICA - 5.3%
- CAJA DE SEGURO SOCIAL - 4.9%
- CAJA DE AHORROS - 3.7%
- MINISTERIO DE SALUD - 2.8%

Distribución Geográfica

Principales Ciudades (Top 5):

- PANAMA - 34.7% de clientes
- ARRAIJAN - 10.3% de clientes
- SAN MIGUELITO - 10.0% de clientes
- LA CHORRERA - 8.9% de clientes
- DAVID - 6.1% de clientes

Demográficos

Distribución de Género:

- Femenino - 78.2% de clientes
- Masculino - 21.8% de clientes

Estado Civil:

- Soltero - 57.0% de clientes
- Casado - 42.9% de clientes

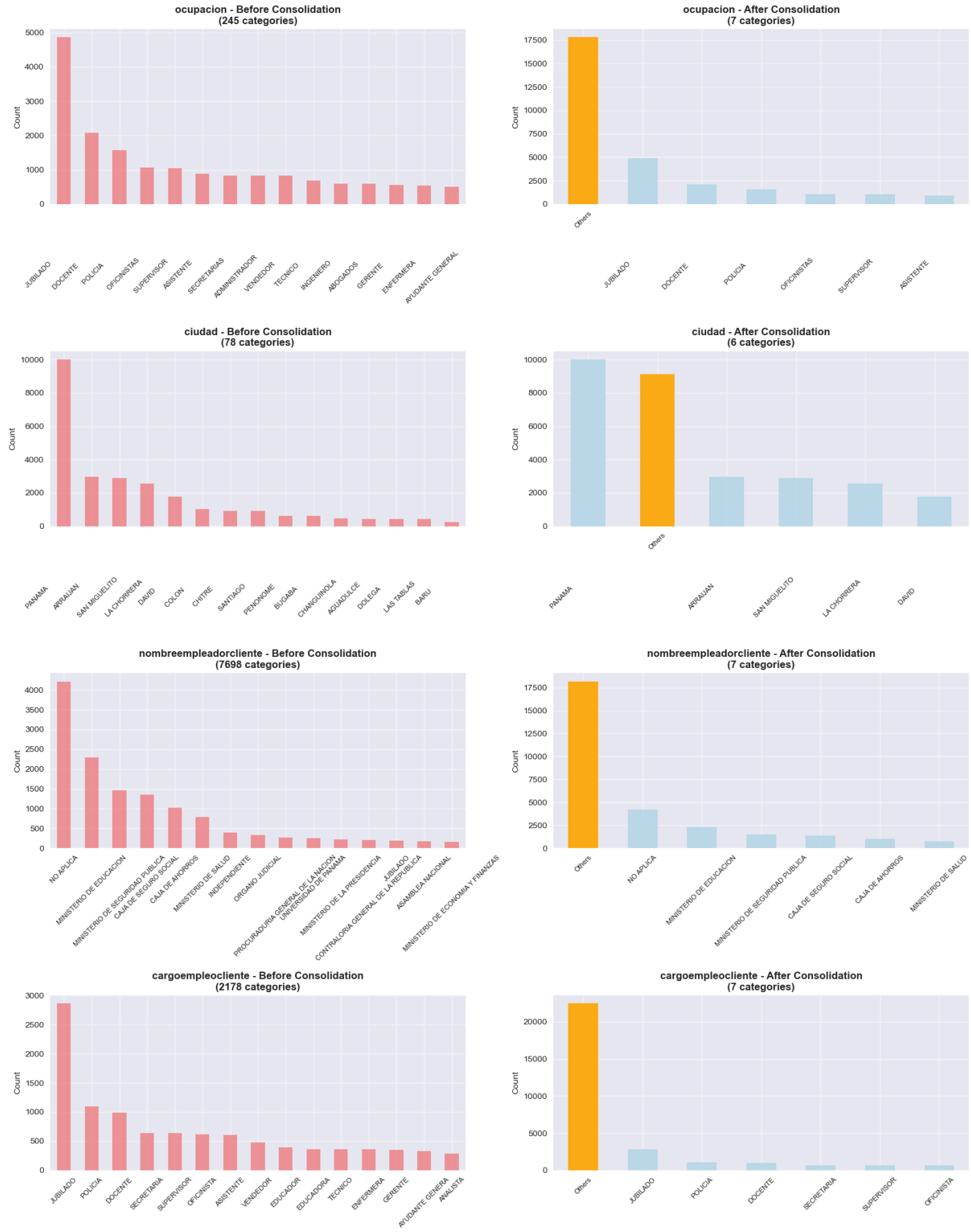

Gráfico 1: antes y después de aplicación de reglas

Estándares de Calidad de Datos

Convenciones Universales de Nombres

Para asegurar consistencia en el registro y procesamiento, se establecieron reglas estándar de nomenclatura:

Tipo de variable	Formato	Ejemplo	Regla de respaldo
Ocupación	MAYÚSCULAS	“JUBILADO”	→ “OTROS”
Empleador	MAYÚSCULAS	“MINISTERIO DE EDUCACION”	→ “OTROS”
Ciudad	MAYÚSCULAS	“PANAMA”	→ “OTROS”
Género	Modo Título	“Femenino”	No aplica
Estado civil	Modo Título	“Soltero”	→ “Otros”

Reglas para Captura de Datos

Para Equipos Operativos:

1. Usar menús desplegables estandarizados (no texto libre)
2. Validación en tiempo real durante registro
3. Seguir ortografía y formato exactamente
4. Mapear valores desconocidos a “Otros” apropiados

Para Integraciones de Sistemas:

1. Normalizar texto antes de almacenar (mayúsculas, espacios, acentos)
2. Validar contra la lista oficial de categorías
3. Notificar entradas inusuales para revisión manual
4. Registrar cambios de categoría en bitácora (log)

Impacto de Negocio y Recomendaciones

Beneficios Inmediatos

1. **Confiabilidad del Modelo:** Menor riesgo de overfitting por variables simplificadas
2. **Eficiencia Operacional:** Reducción del 98.5% en complejidad categórica
3. **Calidad de Datos:** Nombres estandarizados evitan inconsistencias
4. **Escalabilidad:** Codificación apta para valores desconocidos o nuevos

Recomendaciones Estratégicas

Para Operaciones de Negocio:

- Implementar menús desplegables en sistemas de registro

- Capacitar al personal en reglas de nomenclatura
- Monitorear calidad de datos mensualmente
- Crear tablas referenciales de categorías aprobadas

Para Implementación Técnica:

- Desplegar reglas automáticas de validación
- Monitorear cambios en distribución de categorías
- Configurar alertas ante patrones inusuales
- Programar revisiones trimestrales del estándar

Para Mejoras Futuras:

- Agregar nuevas categorías si superan 2% de frecuencia por más de 3 meses
- Evaluar relevancia de categoría emergente
- Medir impacto en desempeño del modelo ante cambios
- Mantener proceso de aprobación por stakeholders
