Reporte de Análisis Exploratorio de Datos

Caja de Ahorros - Proyecto de Predicción de Ingresos

Versión del Documento: 1.0
Fecha: Septiembre 2025
Preparado para: Liderazgo Ejecutivo, Equipo de Ciencia de Datos y Partes Interesadas del Negocio

Resumen Ejecutivo

Este reporte presenta el análisis exploratorio de datos (EDA) integral realizado sobre los datos de clientes para el modelo de predicción de ingresos. Nuestro análisis de **29,319 clientes únicos** reveló insights clave que moldearon nuestra estrategia de modelado y estándares de calidad de datos.

Hallazgos Clave

- **Calidad de Datos:** Consolidación exitosa de 42,549 registros en 29,319 clientes únicos
- **Optimización de Características:** Reducción de complejidad categórica del 98.5% manteniendo relevancia empresarial
- **Logro de Cobertura:** 60-80% de cobertura de datos con características categóricas simplificadas
- **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 primario
Info_Clientes_2.csv	23,502	23,502	Dataset secundario
Final Consolidado	**42,549**	**29,319**	**100%**

Categorías de Características

Nuestro análisis identificó **24 características principales** en cuatro categorías principales:

Demografía del Cliente (6 características)

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

Información de Empleo (4 características)

- Ocupación y posición laboral
- Nombre del empleador y fecha de inicio de empleo

Perfil Financiero (8 características)

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

Características Temporales (6 características)

- Fechas de inicio y fin de cuenta
- Antigüedad en el empleo
- Marcas de tiempo de procesamiento de datos

Desafíos Críticos de Calidad de Datos

Desafío 1: Alta Cardinalidad Categórica

Nuestro análisis inicial reveló cardinalidad extremadamente alta en características categóricas:

Característica	Categorías Originales	Impacto Empresarial
Nombres de Empleadores	7,698 valores únicos	83% aparecen solo una vez
Posiciones Laborales	2,178 valores únicos	72% aparecen solo una vez
Ocupaciones	245 valores únicos	Manejable pero complejo
Ciudades	78 valores únicos	Diversidad geográfica

Implicación Empresarial: Sin manejo adecuado, esto crearía más de 10,000 características del modelo, llevando a:

- Predicciones no confiables debido a datos insuficientes por categoría
- Ineficiencia de memoria y computacional
- Dificultad en interpretación y mantenimiento del modelo

Desafío 2: Inconsistencias en Entrada de Datos

Identificamos problemas comunes de calidad de datos:

- **Variaciones de mayúsculas:** "JUBILADO" vs "jubilado" vs "Jubilado"
- **Caracteres españoles:** "POLICÍA" vs "POLICIA"
- **Sinónimos:** "PROFESOR" vs "DOCENTE" vs "MAESTRO"
- **Problemas de espaciado:** Espacios extra e inconsistencias de formato

---

Solución Estratégica: Consolidación Categórica Inteligente

Nuestro Enfoque: Estrategia "Top-N + Otros"

En lugar de usar métodos de codificación tradicionales que crearían miles de características, implementamos una estrategia de consolidación dirigida por el negocio:

1. **Identificar categorías principales** que proporcionen máximo valor empresarial
2. **Consolidar categorías restantes** en grupos estandarizados de "Otros"
3. **Mantener 60-80% de cobertura de datos** con características simplificadas
4. **Crear reglas de codificación** seguras para producción

Resultados: 98.5% de Reducción de Complejidad

Característica	Antes	Después	Reducción	Cobertura
Nombres de Empleadores	7,698 →	7 categorías	99.9%	60%
Posiciones Laborales	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 Empresariales Aprobadas

Categorías de Empleo

Ocupaciones (Top 6):

- JUBILADO (Jubilado) - 16.6% de clientes
- DOCENTE (Maestros) - 7.1% de clientes
- POLICIA (Policía) - 5.4% de clientes
- OFICINISTAS (Trabajadores de oficina) - 3.7% de clientes
- SUPERVISOR (Supervisores) - 3.6% de clientes
- ASISTENTE (Asistentes) - 3.0% de clientes

Principales Empleadores (Top 6):

- NO APLICA (No aplica/desempleado) - 15.1%
- MINISTERIO DE EDUCACION (Ministerio de Educación) - 8.3%
- MINISTERIO DE SEGURIDAD PUBLICA (Ministerio de Seguridad Pública) - 5.3%
- CAJA DE SEGURO SOCIAL (Caja de Seguro Social) - 4.9%
- CAJA DE AHORROS (Caja de Ahorros) - 3.7%
- MINISTERIO DE SALUD (Ministerio de Salud) - 2.8%

Posiciones Laborales (Top 6):

- JUBILADO (Jubilado)
- POLICIA (Policía)
- DOCENTE (Maestro)
- SUPERVISOR (Supervisor)
- SECRETARIA (Secretaria)
- OFICINISTA (Oficinista)

Distribución Geográfica

Principales Ciudades (Top 5):

- PANAMA (Ciudad de Panamá) - 34.7% de clientes
- ARRAIJAN (Arraiján) - 10.3% de clientes
- SAN MIGUELITO (San Miguelito) - 10.0% de clientes
- LA CHORRERA (La Chorrera) - 8.9% de clientes
- DAVID (David) - 6.1% de clientes

Demografía

Distribución por Género:

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

Estado Civil:

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

Distribución por País:

- PANAMA - 99.9% de clientes

CATEGORÍAS EXACTAS ACEPTADAS - Guía de Referencia para Producción

OCUPACION (Ocupación) - Mantener Top 6

✅ **CATEGORÍAS ACEPTADAS:**

1. **JUBILADO** (Jubilado) - 16.6% de clientes
2. **DOCENTE** (Maestros) - 7.1% de clientes
3. **POLICIA** (Policía) - 5.4% de clientes
4. **OFICINISTAS** (Trabajadores de oficina) - 3.7% de clientes
5. **SUPERVISOR** (Supervisores) - 3.6% de clientes

6. **ASISTENTE** (Asistentes) - 3.0% de clientes

❌ **MARCAR COMO 'OTROS':** Cualquier ocupación NO en la lista anterior

- Ejemplos: "PROFESOR" → "OTROS", "MAESTRO" → "OTROS", "INGENIERO" → "OTROS"

NOMBREEMPLEADORCLIENTE (Empleador) - Mantener Top 6

✅ **CATEGORÍAS ACEPTADAS:**

1. **NO APLICA** (No aplica/desempleado) - 15.1%
2. **MINISTERIO DE EDUCACION** (Ministerio de Educación) - 8.3%
3. **MINISTERIO DE SEGURIDAD PUBLICA** (Ministerio de Seguridad Pública) - 5.3%
4. **CAJA DE SEGURO SOCIAL** (Caja de Seguro Social) - 4.9%
5. **CAJA DE AHORROS** (Caja de Ahorros) - 3.7%
6. **MINISTERIO DE SALUD** (Ministerio de Salud) - 2.8%

❌ **MARCAR COMO 'OTROS':** Cualquier empleador NO en la lista anterior

- Ejemplos: "EMPRESA PRIVADA" → "OTROS", "GOBIERNO" → "OTROS"

CARGOEMPLEOCLIENTE (Posición Laboral) - Mantener Top 6

✅ **CATEGORÍAS ACEPTADAS:**

1. **JUBILADO** (Jubilado)
2. **POLICIA** (Policía)
3. **DOCENTE** (Maestro)
4. **SUPERVISOR** (Supervisor)
5. **SECRETARIA** (Secretaria)
6. **OFICINISTA** (Oficinista)

❌ **MARCAR COMO 'OTROS':** Cualquier posición laboral NO en la lista anterior

- Ejemplos: "GERENTE" → "OTROS", "ANALISTA" → "OTROS"

CIUDAD (Ciudad) - Mantener Top 5

✅ **CATEGORÍAS ACEPTADAS:**

1. **PANAMA** (Ciudad de Panamá) - 34.7% de clientes
2. **ARRAIJAN** (Arraiján) - 10.3% de clientes
3. **SAN MIGUELITO** (San Miguelito) - 10.0% de clientes
4. **LA CHORRERA** (La Chorrera) - 8.9% de clientes
5. **DAVID** (David) - 6.1% de clientes

❌ **MARCAR COMO 'OTROS':** Cualquier ciudad NO en la lista anterior

- Ejemplos: "COLON" → "OTROS", "SANTIAGO" → "OTROS"

SEXO (Género) - Mantener Todas las 2

✅ **CATEGORÍAS ACEPTADAS:**

1. **Femenino** (Femenino) - 78.2% de clientes
2. **Masculino** (Masculino) - 21.8% de clientes

❌ **MARCAR COMO 'Otros':** Cualquier género NO en la lista anterior (casos raros)

ESTADO_CIVIL (Estado Civil) - Mantener Top 2

✅ **CATEGORÍAS ACEPTADAS:**

1. **Soltero** (Soltero) - 57.0% de clientes
2. **Casado** (Casado) - 42.9% de clientes

❌ **MARCAR COMO 'Otros':** Cualquier estado civil NO en la lista anterior

- Ejemplos: "Divorciado" → "Otros", "Viudo" → "Otros"

PAIS (País) - Mantener Top 1

✅ **CATEGORÍAS ACEPTADAS:**

1. **PANAMA** - 99.9% de clientes

❌ **MARCAR COMO 'OTROS':** Cualquier país que NO sea "PANAMA"

- Ejemplos: "COLOMBIA" → "OTROS", "COSTA RICA" → "OTROS"

CONVENCIONES DE NOMENCLATURA CRÍTICAS

Reglas de Sensibilidad a Mayúsculas:

- **Campos en MAYÚSCULAS**: `ocupacion`, `nombreempleadorcliente`, `cargoempleocliente`, `ciudad`, `pais`
  - Usar **"OTROS"** para valores no aprobados
- **Campos en Formato Título**: `sexo`, `estado_civil`
  - Usar **"Otros"** para valores no aprobados

Errores Comunes a Evitar:

```
❌ INCORRECTO → ✅ CORRECTO
"jubilado" → "JUBILADO"
"PROFESOR" → "OTROS" (no aprobado, usar DOCENTE)
"POLICÍA" → "POLICIA" (sin acentos)
"MIN EDUCACION" → "MINISTERIO DE EDUCACION" (nombre completo)
"PANAMA CITY" → "PANAMA"
"Divorciado" → "Otros"
```

Estadísticas Resumen de Consolidación

- **Total de categorías aprobadas**: 29 en todas las características
- **Reducción de complejidad**: 98.5% (de 10,199+ a 29 categorías)
- **Cobertura de datos**: 60-80% con categorías principales
- **Seguridad de producción**: Todos los casos extremos manejados con 'OTROS'/'Otros'

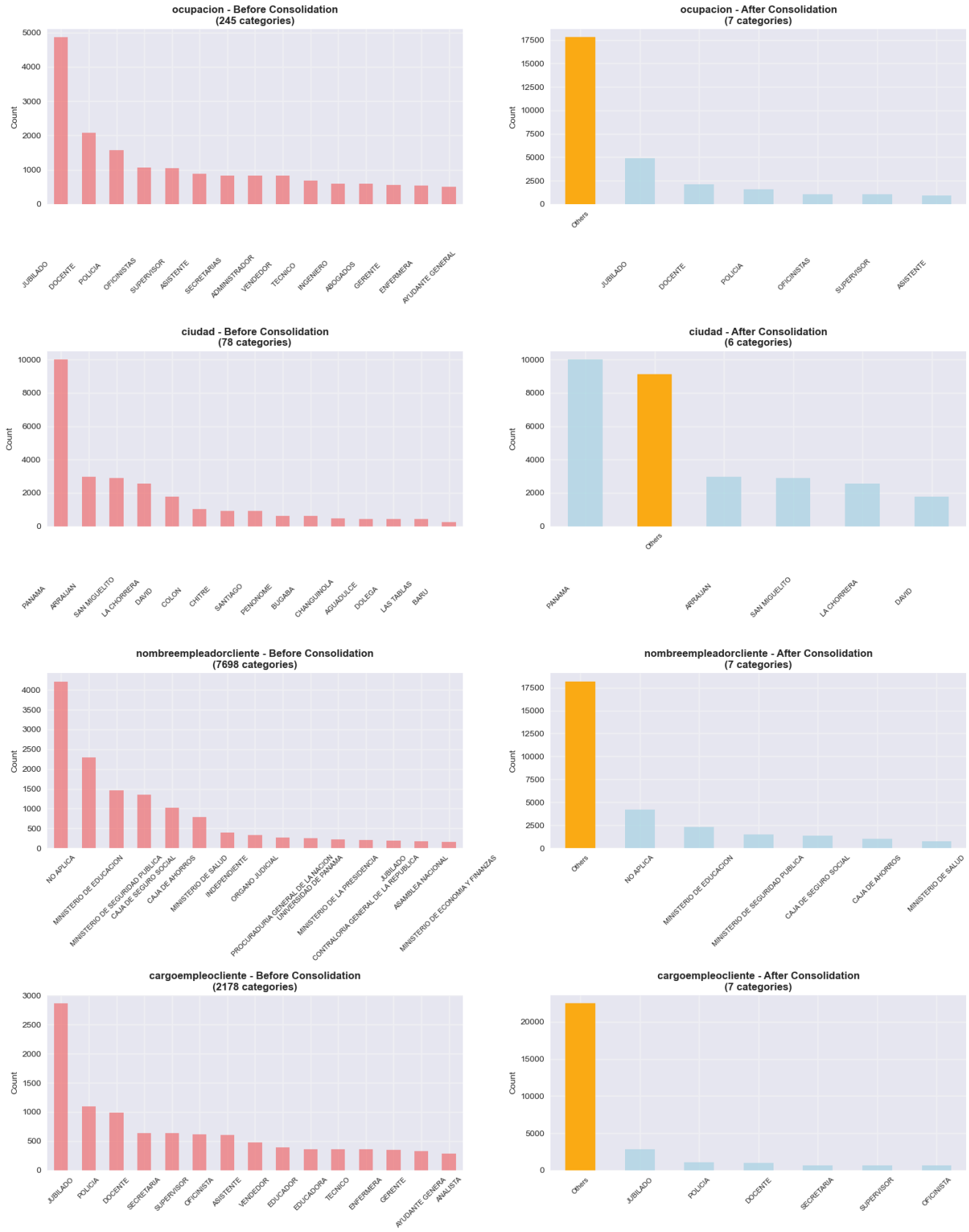

Gráfico 1: Características antes y después de consolidación.

Estándares de Calidad de Datos

Convenciones de Nomenclatura Universal

Para asegurar entrada y procesamiento consistente de datos, establecimos reglas de nomenclatura estandarizadas:

Tipo de Característica	Formato	Ejemplo	Regla de Respaldo
Ocupaciones	TODO EN MAYÚSCULAS	"JUBILADO"	→ "OTROS"
Empleadores	TODO EN MAYÚSCULAS	"MINISTERIO DE EDUCACION"	→ "OTROS"
Ciudades	TODO EN MAYÚSCULAS	"PANAMA"	→ "OTROS"
Género	Formato Título	"Femenino"	No se necesita respaldo
Estado Civil	Formato Título	"Soltero"	→ "Otros"

Directrices de Entrada de Datos

Para Equipos de Operaciones:

1. Usar menús desplegables estandarizados en lugar de texto libre
2. Aplicar validación en tiempo real durante la entrada de datos
3. Seguir reglas exactas de ortografía y formato
4. Mapear categorías desconocidas a grupos apropiados de "Otros"

Para Integración de Sistemas:

1. Normalizar texto antes del almacenamiento (mayúsculas, espacios, acentos)
2. Validar contra listas de categorías aprobadas
3. Marcar entradas inusuales para revisión manual
4. Mantener registros de auditoría de cambios de categorías

Impacto Empresarial y Recomendaciones

Beneficios Inmediatos

1. **Confiabilidad del Modelo:** Reducción del riesgo de sobreajuste a través de características simplificadas
2. **Eficiencia Operacional:** 98.5% de reducción en complejidad categórica
3. **Calidad de Datos:** Convenciones de nomenclatura estandarizadas previenen inconsistencias
4. **Escalabilidad:** Codificación segura para producción maneja valores nuevos/desconocidos

Recomendaciones Estratégicas

Para Operaciones Empresariales:

- Implementar menús desplegables en sistemas de entrada de datos
- Entrenar personal en convenciones de nomenclatura estandarizadas
- Establecer monitoreo mensual de calidad de datos
- Crear tablas de referencia para categorías aprobadas

Para Implementación Técnica:

- Desplegar reglas automáticas de validación de datos
- Monitorear cambios en distribución de categorías a lo largo del tiempo
- Configurar alertas para patrones de datos inusuales
- Programar revisiones trimestrales de estándares de categorías

Para Mejoras Futuras:

- Considerar agregar nuevas categorías si exceden 2% de frecuencia por 3+ meses
- Evaluar relevancia empresarial de categorías emergentes
- Evaluar impacto en rendimiento del modelo de cambios de categorías
- Mantener proceso de aprobación de partes interesadas para modificaciones

---