📊 Exploratory Data Analysis Report

Caja de Ahorros - Income Prediction Project

Document Version: 1.0
Date: September 2025
Prepared for: Executive Leadership, Data Science Team, and Business Stakeholders

Executive Summary

This report presents the comprehensive exploratory data analysis (EDA) conducted on customer data for the income prediction model. Our analysis of **29,319 unique customers** revealed key insights that shaped our modeling strategy and data quality standards.

Key Findings

- **Data Quality:** Successfully consolidated 42,549 records into 29,319 unique customers
- **Feature Optimization:** Reduced categorical complexity by 98.5% while maintaining business relevance
- **Coverage Achievement:** 60-80% data coverage with simplified categorical features
- **Production Readiness:** Established robust data quality standards for operational use

Dataset Overview

Data Sources & Consolidation

Source	Records	Unique Customers	Coverage
Info_Cliente.csv	19,047	19,047	Primary dataset
Info_Clientes_2.csv	23,502	23,502	Secondary dataset
Final Consolidated	**42,549**	**29,319**	**100%**

Feature Categories

Our analysis identified **24 core features** across four main categories:

Customer Demographics (6 features)

- Customer ID and unique identifier
- Age, gender, marital status
- Geographic location (city, country)

Employment Information (4 features)

- Occupation and job position
- Employer name and employment start date

Financial Profile (8 features)

- Account balance and monthly payments
- Loan amounts and interest rates
- Product usage and payment history

Temporal Features (6 features)

- Account start and end dates
- Employment tenure
- Data processing timestamps

Critical Data Quality Challenges

Challenge 1: High Categorical Cardinality

Our initial analysis revealed extremely high cardinality in categorical features:

Feature	Original Categories	Business Impact
Employer Names	7,698 unique values	83% appear only once
Job Positions	2,178 unique values	72% appear only once
Occupations	245 unique values	Manageable but complex
Cities	78 unique values	Geographic diversity

Business Implication: Without proper handling, this would create over 10,000 model features, leading to:

- Unreliable predictions due to insufficient data per category
- Memory and computational inefficiency
- Difficulty in model interpretation and maintenance

Challenge 2: Data Entry Inconsistencies

We identified common data quality issues:

- **Case variations:** "JUBILADO" vs "jubilado" vs "Jubilado"
- **Spanish characters:** "POLICÍA" vs "POLICIA"
- **Synonyms:** "PROFESOR" vs "DOCENTE" vs "MAESTRO"
- **Spacing issues:** Extra spaces and formatting inconsistencies

---

Strategic Solution: Smart Categorical Consolidation

Our Approach: "Top-N + Others" Strategy

Instead of using traditional encoding methods that would create thousands of features, we implemented a business-driven consolidation strategy:

1. **Identify top categories** that provide maximum business value
2. **Consolidate remaining categories** into standardized "Others" groups
3. **Maintain 60-80% data coverage** with simplified features
4. **Create production-safe** encoding rules

Results: 98.5% Complexity Reduction

Feature	Before	After	Reduction	Coverage
Employer Names	7,698 →	7 categories	99.9%	60%
Job Positions	2,178 →	7 categories	99.7%	60%
Occupations	245 →	7 categories	97.1%	39%
Cities	78 →	6 categories	92.3%	80%
Total	**10,199 →**	**29 categories**	**98.5%**	**60-80%**

Approved Business Categories

Employment Categories

Occupations (Top 6):

- JUBILADO (Retired) - 16.6% of customers
- DOCENTE (Teachers) - 7.1% of customers
- POLICIA (Police) - 5.4% of customers
- OFICINISTAS (Office workers) - 3.7% of customers
- SUPERVISOR (Supervisors) - 3.6% of customers
- ASISTENTE (Assistants) - 3.0% of customers

Major Employers (Top 6):

- NO APLICA (Not applicable/unemployed) - 15.1%
- MINISTERIO DE EDUCACION (Ministry of Education) - 8.3%
- MINISTERIO DE SEGURIDAD PUBLICA (Ministry of Public Security) - 5.3%
- CAJA DE SEGURO SOCIAL (Social Security Fund) - 4.9%
- CAJA DE AHORROS (Savings Bank) - 3.7%
- MINISTERIO DE SALUD (Ministry of Health) - 2.8%

Geographic Distribution

Major Cities (Top 5):

- PANAMA (Panama City) - 34.7% of customers
- ARRAIJAN (Arraiján) - 10.3% of customers
- SAN MIGUELITO (San Miguelito) - 10.0% of customers
- LA CHORRERA (La Chorrera) - 8.9% of customers
- DAVID (David) - 6.1% of customers

Demographics

Gender Distribution:

- Femenino (Female) - 78.2% of customers
- Masculino (Male) - 21.8% of customers

Marital Status:

- Soltero (Single) - 57.0% of customers
- Casado (Married) - 42.9% of customers

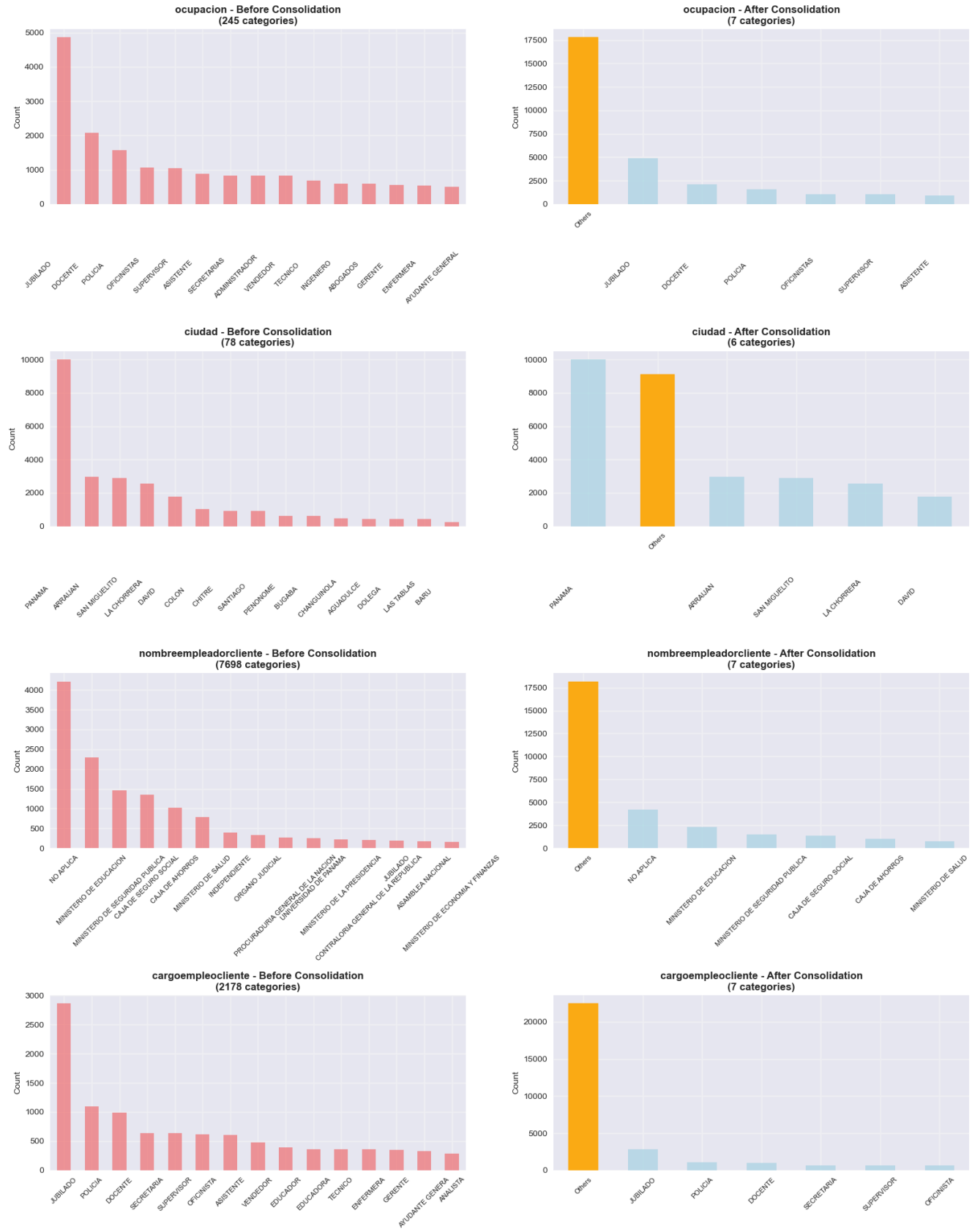

Gráfico 1: Features antes y después de consolidación.

Data Quality Standards

Universal Naming Conventions

To ensure consistent data entry and processing, we established standardized naming rules:

Feature Type	Format	Example	Fallback Rule
Occupations	ALL UPPERCASE	"JUBILADO"	→ "OTROS"
Employers	ALL UPPERCASE	"MINISTERIO DE EDUCACION"	→ "OTROS"
Cities	ALL UPPERCASE	"PANAMA"	→ "OTROS"
Gender	Title Case	"Femenino"	No fallback needed
Marital Status	Title Case	"Soltero"	→ "Otros"

Data Entry Guidelines

For Operations Teams:

1. Use standardized dropdown menus instead of free text
2. Apply real-time validation during data entry
3. Follow exact spelling and formatting rules
4. Map unknown categories to appropriate "Others" groups

For System Integration:

1. Normalize text before storage (case, spacing, accents)
2. Validate against approved category lists
3. Flag unusual entries for manual review
4. Maintain audit logs of category changes

Business Impact & Recommendations

Immediate Benefits

1. **Model Reliability:** Reduced overfitting risk through simplified features
2. **Operational Efficiency:** 98.5% reduction in categorical complexity
3. **Data Quality:** Standardized naming conventions prevent inconsistencies
4. **Scalability:** Production-safe encoding handles new/unknown values

Strategic Recommendations

For Business Operations:

- Implement dropdown menus in data entry systems
- Train staff on standardized naming conventions
- Establish monthly data quality monitoring
- Create reference tables for approved categories

For Technical Implementation:

- Deploy automated data validation rules
- Monitor category distribution changes over time
- Set up alerts for unusual data patterns
- Schedule quarterly reviews of category standards

For Future Enhancements:

- Consider adding new categories if they exceed 2% frequency for 3+ months
- Evaluate business relevance of emerging categories
- Assess model performance impact of category changes
- Maintain stakeholder approval process for modifications
