Examen Final
82.05 - Análisis Predictivo
Catalina Trevisan (64990)

Predicción de default de préstamos

Sistema de Scoring Automático para Evaluación crediticia

Caso de negocio:

Las instituciones financieras enfrentan pérdida significativas (del 2-5% de su cartera crediticia) por defaults, como consecuencia de un proceso manual ineficiente:

- 3-7 días de procesamiento
- Evaluaciones subjetivas (diferentes analistas → diferentes decisiones)
- Alto costo operativo de evaluación
- ~15% de defaults no detectados

SOLUCIÓN PROPUESTA

Sistema de scoring automático basado en ML que:

- Evalúa en tiempo real (<5 minutos)
- Objetivo y consistente (mismo algoritmo, mismos criterios)
- Optimiza ROI (balance entre aprobaciones y pérdidas)
- Escalable (miles de evaluaciones simultáneas)

IMPACTO ESPERADO

Reducción de defaults
Reducción del tiempo procesamiento
Optimización del ROI
Mejora en la satisfacción cliente

Dataset:

Fuente: Loan Approval | Classification - Kaggle

Dimensiones: 44,993 préstamos × **14** variables

Target: loan_status (1=Default, 0=Paid)

Balance: 22.2% defaults vs **77.8%** pagados

Calidad: Sin valores faltantes/duplicados, outliers identificados

FEATURES CLAVE:

DEMOGRÁFICAS:

- person_age , person_gender , person_education
- person_income , person_emp_exp , person_home_ownership

DEL PRÉSTAMO:

- loan_amnt , loan_intent , loan_int_rate
- loan_percent_income

CREDITICIAS:

- cb_person_cred_hist_length , credit_score
- previous_loan_defaults_on_file

Análisis exploratorio

Factores de Riesgo DEMOGRÁFICO

- **Edad +55:** +24.00% de default rate
- **Educación secundaria:** 22.31% default rate
- **Tipo vivienda OTHER:** 33.33% default rate

Factores de Riesgo del PRÉSTAMO

- **Propósito DEBTCONSOLIDATION:** 30.27% default
- **Tasa de interés**: 12.86% (default) vs 10.48% (paid)

CORRELACIONES con default

Positivas (riesgo):

- `loan_percent_income` : +0.38
- `loan_int_rate` : +0.33
- `loan_amnt` : +0.11

Negativas (protección):

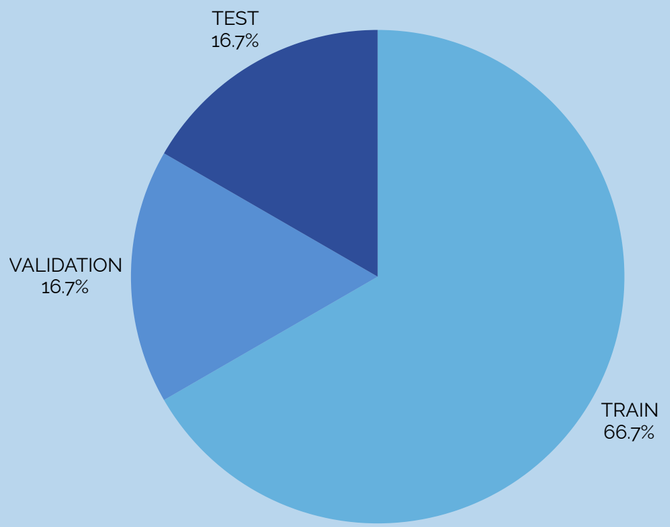
- `person_income` : -0.17
- `credit_score` : -0.01

Evaluación y partición de datos

Métrica de Evaluación: ROC-AUC

- Apropiada para clasificación binaria con desbalance (22.2% defaults)
- Mide capacidad de discriminar entre clases
- Independiente del umbral de clasificación
- Permite comparar modelos objetivamente

Partición de datos



Estratificación: Se mantiene la proporción 22.2% defaults en cada set

Modelo Baseline

BASELINE 1

Most Frequent (Predicción por Mayoría)

- Predice siempre la clase mayoritaria (Paid)
- ROC-AUC: ~0.5000

BASELINE 2

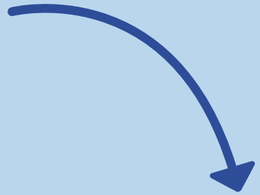
Stratified (Predicción Estratificada)

- Predice según distribución de clases
- ROC-AUC: ~0.5049

Cualquier modelo debe superar ROC-AUC > 50

Selección de modelos

Modelo	ROC-AUC	F1-Score	Precision	Recall
XGBoost	0.9739	0.8229	0.8684	0.7820
Random Forest	0.9694	0.8049	0.8844	0.7385
Gradient Boosting	0.9675	0.7961	0.8697	0.7345
Logistic Regression	0.9485	0.7310	0.6989	0.9145
Decision Tree	0.9417	0.7567	0.6668	0.8745



Modelo ganador (mayor ROC-AUC en validation set)

Modelo final y optimización

HIPERPARÁMETROS OPTIMIZADOS

Método: Grid Search con Cross-Validation (5-fold)

Mejores parámetros:

- n-estimators: 300
- max_depth: 5
- learning_rate: 0.01
- subsample: 1.0

Mejor ROC-AUC (CV): 0.9767

FEATURES ENGINEERED

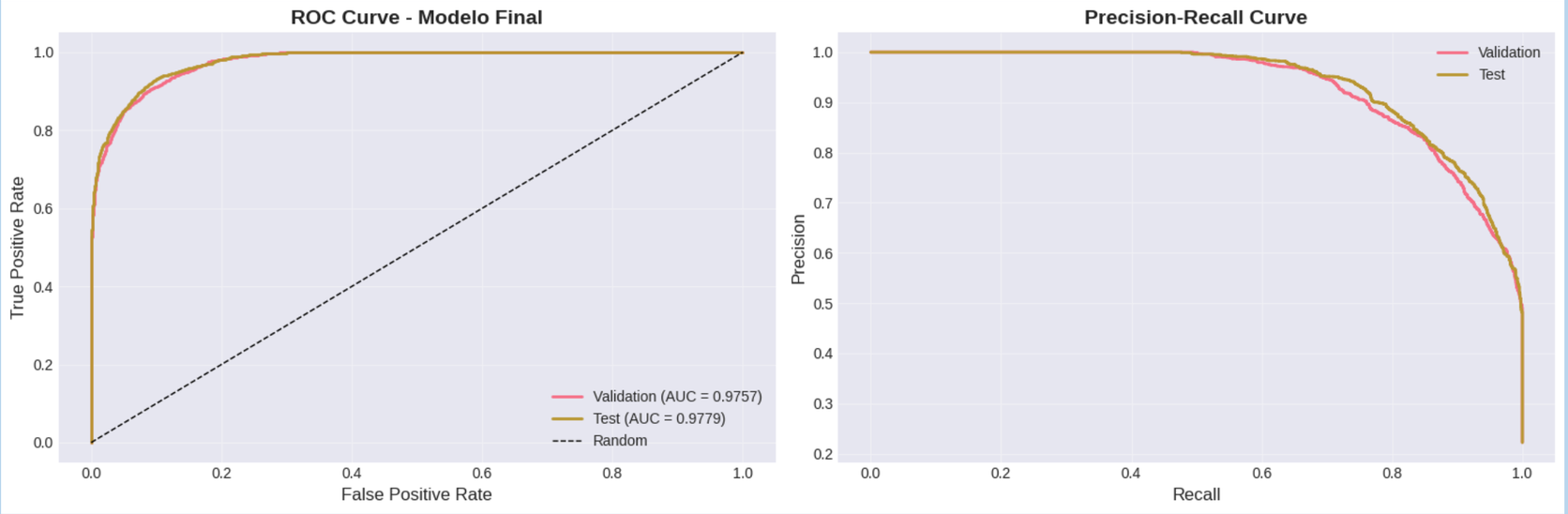
income_to_loan_ratio
debt_burden
credit_to_income
exp_to_age_ratio
Binning de edad, ingreso y credit score

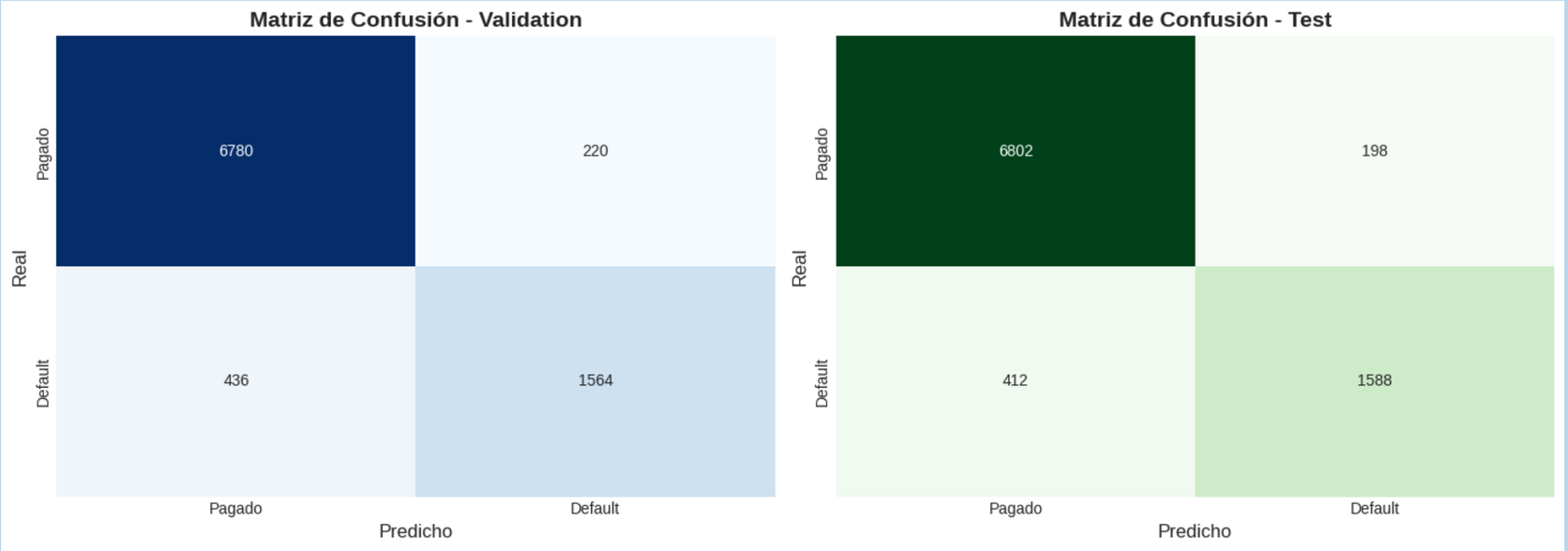
Resultados finales

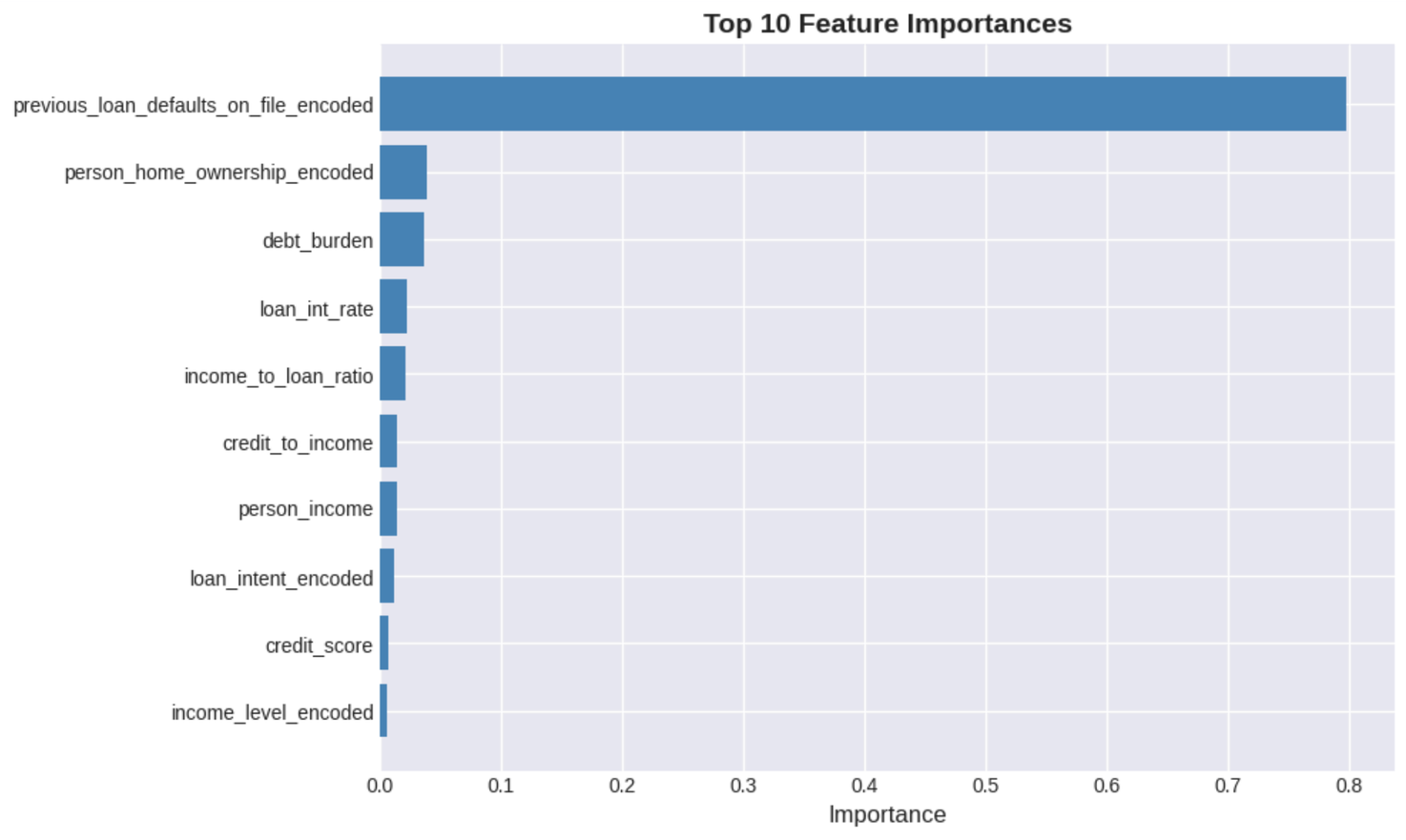
	Validation	Test
ROC-AUC	0.9757	0.9779
F1 Score	0.8266	0.8389
Precision	0.8767	0.8891
Recall	0.7820	0.7940

MEJORA RESPECTO A BASELINE

- Baseline ROC-AUC: 0.505
- Modelo Final ROC-AUC: 0.9767
- Mejora absoluta: 0.4171
- Mejora relativa: 93.4%


ROC Curve - Modelo Final
True Positive Rate
False Positive Rate
Validation (AUC = 0.9757)
Test (AUC = 0.9779)
Random
Precision-Recall Curve
Precision
Recall
Validation
Test


Matriz de Confusión - Validation
Real
Predicho
Pagado
Default
6780
220
436
1564
Matriz de Confusión - Test
6802
198
412
1588


Top 10 Feature Importances
previous_loan_defaults_on_file_encoded
person_home_ownership_encoded
debt_burden
loan_int_rate
income_to_loan_ratio
credit_to_income
person_income
loan_intent_encoded
credit_score
income_level_encoded
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Importance

Limitaciones y mejoras futuras

LIMITACIONES ACTUALES

1. **Desbalance de clases**: 22.2% vs 77.8% - puede afectar recall en minoría
2. **Temporal**: Sin validación en diferentes períodos de tiempo
3. **Features limitadas**: No incluye variables macroeconómicas
4. **Interpretabilidad**: Modelos tree-based menos interpretables que regresión

MEJORAS FUTURAS

1. **Técnicas de balanceo**: SMOTE, ADASYN para mejorar detección de defaults
2. **Ensemble avanzado**: Stacking o Voting Classifier combinando varios modelos
3. **Calibración**: Ajustar probabilidades con CalibratedClassifierCV
4. **Explicabilidad**: Implementar SHAP values para interpretación