# Predicción de Éxito Comercial en Películas

Lautaro Calaza

Examen Final - Análisis Predictivo

# Problema y Contexto

## Pregunta de Negocio

- ¿Cómo predecir si una película será exitosa comercialmente?
- ¿Qué factores determinan el éxito?
- Optimizar decisiones de inversión

## Dataset

- 17,183 películas con revenue > 0
- Desbalanceado: 75% no exitosas, 25% exitosas
- Features: presupuesto, géneros, idiomas, métricas de audiencia

# Metodología

01

## Exploración

Análisis detallado de variables, distribuciones, correlaciones

02

## Feature Engineering

50 features a partir de fechas, géneros, idiomas, presupuesto

03

## Preprocesamiento

Winsorizing P1–P99, imputación, StandardScaler

04

## Modelado

Comparación de 5 algoritmos (Logistic Regression, Ridge, Lasso, Random Forest, XGBoost)

05

## Optimización

GridSearchCV + threshold tuning

# Comparación de Modelos

*Modelo ganador:* XGBoost con F1 = 0.7951 (+2.75% sobre baseline)

| Modelo | F1-Score | Precision | Recall | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.7740 | 0.7097 | 0.8510 | 0.8758 |
| Ridge | 0.7567 | 0.8303 | 0.6950 | 0.8883 |
| Lasso | 0.7415 | 0.8699 | 0.6461 | 0.8874 |
| Random Forest | 0.7881 | 0.7492 | 0.8312 | 0.8883 |
| **XGBoost (Optimizado)** | **0.7951** | **0.7620** | **0.8312** | **0.8929** |

# Por qué XGBoost ganó

### Ridge y Lasso: Problema de Conservadurismo

- Ridge: Precision 83% (muy conservador) pero Recall 69.5% (pierde muchas oportunidades)
- Lasso: Precision 87% (extremadamente conservador) pero Recall 64.6% (detecta pocas películas exitosas)

### Random Forest: Buen balance pero insuficiente

- F1: 0.7881 (1.4% menos que XGBoost)
- Recall similar pero menor Precision

### XGBoost: Mejor balance global

- F1: 0.7951 (métrica elegida para desbalanceo)
- Precision 76% y Recall 83% - ambas sólidas
- Ganador claro en desempeño y generalización

# Métricas del Modelo Final

| Métrica | Valor |
| --- | --- |
| F1-Score | 0.7951 |
| Precision | 0.7620 |
| Recall | 0.8312 |
| ROC-AUC | 0.9535 |

De 100 películas que el modelo predice como exitosas, 76 realmente lo son. Del 100% de películas exitosas reales, el modelo detecta 83%. Excelente discriminación entre clases.

# Top 5 Features Más Importantes

1 budget_log (14.04%)

Presupuesto en escala logarítmica - Predictor dominante

2 vote_count (10.26%)

Cantidad de votos/reviews - Indicador de popularidad

3 budget (5.11%)

Presupuesto original - Correlación con éxito

4 lang_zh (3.44%)

Películas en idioma chino - Mercado emergente

5 genre_comedy (2.41%)

Género comedia - Mayor probabilidad de éxito

# Insights Clave

El presupuesto es el predictor dominante de éxito comercial

La cantidad de reviews/votos correlaciona fuertemente con éxito

Ciertos idiomas (chino, francés, coreano) tienen mayor probabilidad de éxito

Géneros como comedia y familia tienden a ser más exitosos

Dataset desbalanceado requirió uso de F1-Score como métrica principal

Threshold óptimo: 0.55 (maximiza balance entre precision y recall)

# Análisis de Errores

## Matriz de Confusión

| | PRED. NO EXITOSA | PRED. EXITOSA |
|---|---|---|
| Real No Exitosa | 2330 | 248 |
| Real Exitosa | 137 | 722 |

TP: 722 | TN: 2330

FP: 248 | FN: 137

## Interpretación

- **TP (722):** Correctamente predichas como exitosas
- **TN (2330):** Correctamente predichas como no exitosas
- **FP (248):** Falsos positivos (invertir en fracasos)
- **FN (137):** Falsos negativos (perder oportunidades)

# Hiperparámetros Optimizados

## Parámetros XGBoost

- **max_depth:** 7 - Profundidad de árboles
- **learning_rate:** 0.1 - Velocidad de aprendizaje
- **subsample:** 0.8 - Fracción de muestras por árbol
- **colsample_bytree:** 0.7 - Fracción de features por árbol
- **n_estimators:** 100 - Cantidad de árboles
- **threshold:** 0.55 - Punto de corte optimizado

## Proceso de Optimización

GridSearchCV con 324 combinaciones de parámetros.
Métrica: F1-Score. Validación cruzada: 5 folds

# Deliverables del Proyecto

## Código y Modelos

- train_model.py - Script completo de entrenamiento
- predict_model.py - Script para hacer predicciones
- xgb_model_final.pkl - Modelo entrenado (340 KB)
- scaler_final.pkl - StandardScaler para normalización
- model_metadata.pkl - Metadatos del modelo

## Documentación

- README.md - Documentación completa del proyecto
- requirements.txt - Dependencias Python
- GitHub Repository - Código versionado y accesible

# Limitaciones del Modelo

Entrenado solo con películas que tienen revenue > 0 (sesgo de supervivencia)

Features basadas principalmente en metadatos y presupuesto

No incluye datos de redes sociales ni tendencias de mercado

No considera factores externos (cambios económicos, pandemias, etc.)

Target definido como percentil 75 de revenue (puede variar según contexto)

# Conclusiones

### Modelo Exitoso

XGBoost con F1 = 0.7951 proporciona predicciones confiables de éxito comercial en películas

### Aplicación Práctica

El modelo puede asistir en decisiones de inversión, identificando películas con alta probabilidad de éxito comercial

### Próximos Pasos

Integración en pipeline de análisis, reentrenamiento periódico con nuevos datos, análisis de nuevos features