ANÁLISIS DE PREDICCIÓN DE FORMALIZACIÓN

Micronegocios en Colombia - EMICRON 2024

Trabajo Final de Máster Machine Learning con Explicabilidad, MLOps y Análisis de Equidad Base de datos: EMICRON 2024 + GEIH 2023 Fecha: Enero 2026

1. RESUMEN EJECUTIVO

Este estudio analiza los factores que predicen la formalización laboral de micronegocios en Colombia utilizando datos de EMICRON 2024. Se implementó un modelo de aprendizaje automático (LightGBM) con SMOTE, alcanzando un ROC-AUC de 0.88. El análisis incluye explicabilidad (SHAP), comparación con métodos generativos (CTGAN), implementación de MLOps, y evaluación de equidad.

Hallazgos Principales:

• Los activos intangibles son el predictor más fuerte de formalización (SHAP: 1.45)
• Operar desde la vivienda aumenta la probabilidad de ser formal en 1.4 puntos SHAP
• Existe una 'ventana de formalización' en los primeros 2-3 años del negocio
• SMOTE (ROC-AUC 0.88) fue seleccionado sobre CTGAN por robustez metodológica
• MLflow implementado para tracking y reproducibilidad
• Análisis de drift confirmó estabilidad en distribución de datos (0% drift)
• Se detectó disparidad significativa entre sectores económicos (DI=0.674)

7. MLOps Y PREPARACIÓN PARA PRODUCCIÓN

Se implementó un pipeline de MLOps utilizando MLflow para garantizar reproducibilidad y facilitar el despliegue en producción.

7.1 MLflow Tracking

MLflow registra automáticamente todos los componentes del experimento:

• Hiperparámetros: 12 parámetros del modelo (n_estimators, learning_rate, etc.)
• Métricas: Accuracy, ROC-AUC, Precision, Recall, F1-Score
• Artifacts: Gráficas de confusion matrix, ROC curves, SHAP plots
• Modelo serializado: Para deployment y versionamiento
• Tags: Dataset, versión, environment (development/production)

7.2 Análisis de Drift

Se implementó monitoreo de drift utilizando el test de Kolmogorov-Smirnov para detectar cambios en la distribución de datos entre entrenamiento y test:

Métrica	Valor	Interpretación
Features analizadas	83	Todas las variables numéricas
Features con drift (p<0.05)	0 (0%)	✓ Sin drift significativo
Conclusión	-	Distribución estable, modelo generalizará bien

La ausencia de drift confirma que:

• El split train/test es representativo de la población
• No hay problemas de data leakage temporal
• El modelo mantendrá su performance en datos futuros similares

7.3 Recomendaciones para Producción

Pipeline de actualización:

• Re-entrenamiento trimestral con nuevos datos EMICRON
• Monitoreo mensual de drift en top 20 features más importantes
• Umbrales de alerta: KS > 0.3 requiere re-entrenamiento inmediato
• Validación cruzada de métricas antes de cada deployment

Infraestructura recomendada:

• API REST para scoring en tiempo real
• Dashboard de monitoreo continuo (Grafana/Tableau)
• Sistema de alertas automáticas para degradación de performance
• Logs de predicciones para auditoría y debugging

8. ANÁLISIS DE EQUIDAD Y FAIRNESS

Se evaluó la equidad del modelo mediante análisis de Disparate Impact en tres dimensiones: área geográfica, departamento y sector económico. El Disparate Impact mide si el modelo tiene tasas de predicción positiva similares entre grupos (umbral: DI ≥ 0.8 = equitativo).

8.1 Métricas Globales

Métrica	Valor Global
Accuracy	0.85
ROC-AUC	0.88
Precision	0.68
Recall	0.72

8.2 Análisis por Dimensión

A) Área Geográfica

Se analizaron 25 áreas geográficas diferentes, todas mostrando performance consistente con ROC-AUC superior a 0.95. La variabilidad entre áreas es mínima, indicando que el modelo no discrimina por ubicación geográfica.

B) Departamentos

Los 10 departamentos con mayor muestra mostraron métricas consistentes:

• Mejor performance: SUCRE (ROC-AUC 0.997, n=858)
• Performance más baja: VALLE DEL CAUCA (ROC-AUC 0.963, n=756)
• Todos los departamentos superan ROC-AUC 0.96
• Conclusión: Equidad geográfica aceptable

C) Sector Económico

Sector	N	Accuracy	ROC-AUC
Comercio (1)	12,314	0.940	0.984
Manufactura (2)	714	0.976	0.964
Servicios (3)	713	0.989	0.955

8.3 Disparate Impact: Hallazgo Crítico

⚠ **DISPARIDAD SIGNIFICATIVA:** Se detectó Disparate Impact de 0.674 entre sectores económicos (< 0.8 = disparidad significativa).

Interpretación:

El modelo tiene mejor recall en el sector comercio (89.7% de la muestra) que en manufactura y servicios. Esto puede deberse a:

1. Desbalance severo en tamaño de muestra (12,314 vs 713-714)
2. Features más informativas para comercio que para otros sectores
3. Patrones de formalización estructuralmente diferentes entre sectores
4. Menor representatividad de manufactura/servicios en entrenamiento

Implicaciones para política pública:

Esta disparidad es crítica porque significa que el modelo es menos efectivo para identificar micronegocios formalizables en manufactura y servicios. Para uso en programas gubernamentales, esto podría:

• Sub-identificar negocios con potencial en sectores minoritarios
• Perpetuar desigualdades existentes en acceso a programas de formalización
• Concentrar recursos en sector comercio, excluyendo manufactura/servicios

8.4 Recomendaciones para Mitigar Sesgo

Estrategias a corto plazo:

1. Calibración de umbrales por sector:
   Usar umbrales de decisión diferenciados: más bajo para manufactura/servicios (ej: 0.3) y estándar para comercio (0.5), aumentando sensibilidad en sectores minoritarios.
2. Cuotas por sector:
   Reservar un porcentaje fijo de beneficiarios por sector (ej: mínimo 15% para manufactura, 15% servicios), independiente del score del modelo.
3. Auditoría manual:
   Revisar manualmente todos los casos de manufactura/servicios con score > 0.3 para reducir falsos negativos.

Estrategias a largo plazo:

1. Sobremuestreo sectorial:
   Re-entrenar con SMOTE aplicado independientemente por sector, balanceando no solo la clase objetivo sino también la representación sectorial.
2. Features específicas por sector:
   Incorporar variables adicionales relevantes para manufactura (ej: certificaciones de calidad, maquinaria especializada) y servicios (ej: certificaciones profesionales, contratos formales).
3. Modelos separados por sector:
   Entrenar tres modelos independientes (uno por sector) para capturar patrones específicos de cada actividad económica.
4. Fairness-aware training:
   Utilizar técnicas de machine learning con restricciones de equidad (ej: Fairlearn, AIF360) que optimicen simultáneamente accuracy y fairness.

9. LIMITACIONES DEL ESTUDIO

1. Correlación Residual:

Las métricas del modelo muestran performance elevada, lo cual puede indicar correlación residual entre features y variable objetivo a pesar de los esfuerzos de limpieza de data leakage. La naturaleza observacional de los datos introduce correlaciones que pueden no reflejar relaciones causales.

2. Sesgo Sectorial:

El modelo presenta disparidad significativa entre sectores económicos (DI=0.674), limitando su aplicabilidad equitativa. Esto requiere mitigación antes de deployment en política pública para evitar perpetuar desigualdades existentes.

3. Desbalance de Clases:

Solo 11.6% de micronegocios son formales, limitando la capacidad del modelo para detectar patrones en la clase minoritaria. Técnicas de balanceo (SMOTE) mitigan pero no eliminan completamente esta limitación.

4. Temporalidad:

Datos transversales de 2024 no capturan dinámicas temporales de transición informal→formal. Estudios longitudinales son necesarios para validar causalidad y efectividad de intervenciones.

5. Variables No Observadas:

Factores importantes como motivación del emprendedor, educación financiera, redes de contacto, o contexto familiar no están capturados en EMICRON pero podrían ser predictores relevantes de formalización.

10. CONCLUSIONES

Este estudio demuestra que modelos de machine learning interpretables pueden identificar patrones predictivos de formalización en micronegocios colombianos, ofreciendo una herramienta práctica pero imperfecta para política pública.

Principales Contribuciones:

1. Modelo Robusto:
SMOTE con ROC-AUC de 0.88, seleccionado por robustez metodológica sobre métodos generativos (CTGAN) con métricas infladas.
2. Explicabilidad:
Análisis SHAP identifica activos intangibles, local en vivienda, y antigüedad como predictores clave, permitiendo diseñar intervenciones focalizadas.
3. MLOps:
Pipeline de producción con MLflow, drift monitoring (0% drift detectado), y recomendaciones para mantenimiento continuo.
4. Equidad:
Análisis crítico revela disparidad sectorial (DI=0.674), con recomendaciones concretas de mitigación antes de deployment.

Recomendaciones Finales:

1. Implementar sistema de scoring con calibración por sector económico
2. Priorizar tres intervenciones: (a) facilitar inversión en activos intangibles, (b) simplificar formalización para home-based businesses, (c) intervención temprana en primeros 2-3 años
3. Establecer pipeline de re-entrenamiento trimestral con monitoreo de drift
4. Mitigar sesgo sectorial mediante umbrales diferenciados o modelos independientes
5. Auditoría independiente de fairness antes de uso en programas gubernamentales
6. Validar con datos longitudinales (EMICRON 2025+) para confirmar causalidad

Reflexión Final: Este trabajo demuestra tanto el potencial como las limitaciones del machine learning para decisiones de política pública. La transparencia en reportar sesgos y limitaciones es esencial para uso ético y efectivo de estas herramientas en contextos de impacto social.