

Reporte de Desarrollo de Modelo

Caja de Ahorros - Sistema de Predicción de Ingresos

Versión del Documento: 1.0

Fecha: Septiembre 2025

Preparado para: Liderazgo Ejecutivo, Equipo de Ciencia de Datos y Stakeholders de Negocio

Resumen Ejecutivo

Este reporte documenta el proceso integral de desarrollo de modelo para predecir ingresos de clientes en Caja de Ahorros. Nuestro análisis de **28,665 clientes** resultó en un sistema robusto de machine learning capaz de predicciones precisas de ingresos con manejo apropiado de casos extremos y requerimientos de negocio.

Logros Clave

- Ingeniería de Características:** Desarrollo de 22 características predictivas a partir de datos brutos de clientes
- Calidad de Datos:** Implementación de pipeline robusto de preprocesamiento manejando valores faltantes y outliers
- Análisis de Distribución de Ingresos:** Comprensión integral de patrones de ingresos de clientes
- Preparación para Producción:** Pipeline escalable de preprocesamiento para despliegue operacional

Preparación de Dataset e Ingeniería de Características

Conjunto Final de Características

Nuestro dataset de modelado incluye **22 características cuidadosamente diseñadas** en cuatro categorías:

Demografía de Clientes (5 características)

- Indicadores de edad y demográficos
- Codificación geográfica (ciudad, país)
- Clasificaciones de estado civil y género

Perfil de Empleo y Financiero (8 características)

- Codificación por frecuencia de ocupación y empleador
- Balance de cuenta y montos de pago
- Montos de préstamos y tasas de interés
- Cálculos de antigüedad laboral

Características Temporales (6 características)

- Fecha de inicio de empleo (días desde referencia)
- Fecha de apertura de cuenta (días desde referencia)
- Duración de contrato y métricas de antigüedad

Indicadores Diseñados (3 características)

- Banderas de valores faltantes para campos críticos
- Ratios préstamo-a-pago
- Puntuaciones de estabilidad profesional

Pipeline de Preprocesamiento de Datos

Nuestro sistema de preprocesamiento maneja desafíos de datos del mundo real:

Proceso	Descripción	Impacto Empresarial
Manejo de Valores Faltantes	Imputación por mediana con indicadores de valores faltantes	Preserva información mientras permite predicciones
Conversión de Fechas	Convertir fechas a días desde referencia	Permite reconocimiento de patrones temporales
Codificación Categórica	Codificación por frecuencia para características de alta cardinalidad	Mantiene poder predictivo con eficiencia
Creación de Características	Ratios de préstamos e indicadores de estabilidad	Captura relaciones relevantes para el negocio

Análisis de Distribución de Ingresos

Estadísticas Generales de Ingresos

Nuestro análisis reveló patrones importantes en la distribución de ingresos de clientes:

Métrica	Valor	Insight de Negocio
Total de Clientes	28,665	Dataset completo después de filtrado de calidad
Ingreso Promedio	\$1,494.28	Nivel promedio de ganancias de clientes
Ingreso Mediano	\$1,194.00	Ingreso típico de cliente (menos afectado por outliers)
Rango de Ingresos	\$0.01 - \$5,699.89	Amplio rango requiere modelado robusto
Desviación Estándar	\$1,095.34	Variabilidad significativa de ingresos

Insights de Distribución de Ingresos

Cuartiles de Ingresos:

- **Percentil 25:** \$750.00 (Clientes de ingresos bajos)
- **Percentil 50:** \$1,194.00 (Ingreso mediano)
- **Percentil 75:** \$1,912.86 (Clientes de ingresos altos)
- **Percentil 95:** \$3,827.54 (Principales generadores de ingresos)

Análisis de Segmentos Especiales de Ingresos

Segmento de Ingresos Bajos (< \$500)

Hallazgos Clave:

- **Cantidad:** 1,388 clientes (4.84% del total)
- **Ingreso Promedio:** \$269.09
- **Rango de Ingresos:** \$0.01 - \$499.52

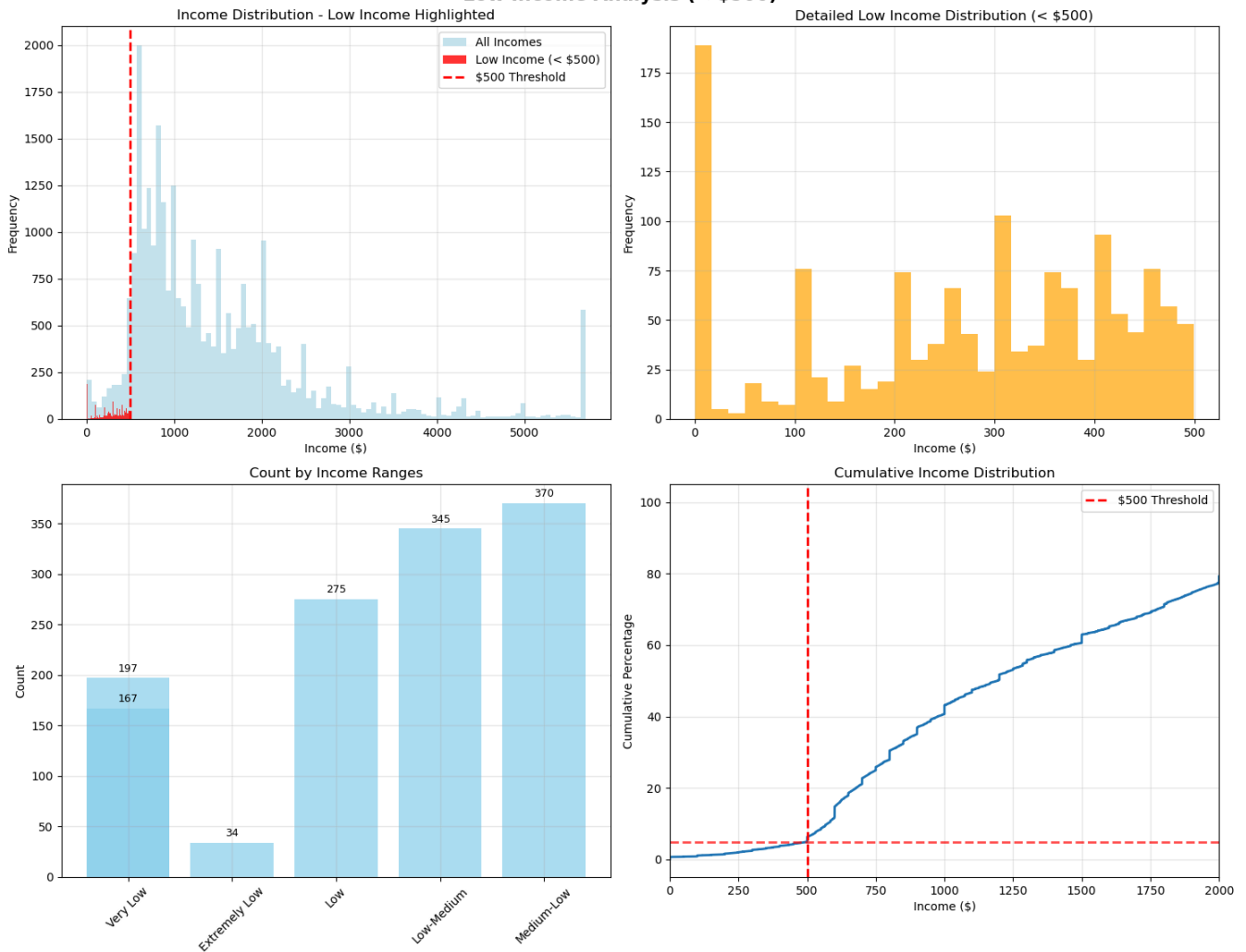
Características:

- Pagos mensuales menores (\$64.17 vs \$132.65 promedio)
- Montos de préstamo mayores (\$13,056.68 vs \$3,508.43 promedio)
- Demografía ligeramente mayor (49.93 vs 48.84 años promedio)

Implicaciones para Modelado:

- Requiere métricas de evaluación robustas
- Puede beneficiarse de funciones de pérdida ponderadas
- Necesita monitoreo cuidadoso para precisión de predicción

Low Income Analysis (< \$500)



[Gráfico 1: Low-income Distribution]

Segmento de Ingresos Altos (> \$5,000)

Hallazgos Clave:

- **Cantidad:** 747 clientes (2.61% del total)
- **Ingreso Promedio:** \$5,618.99
- **Rango de Ingresos:** \$5,008.00 - \$5,699.89

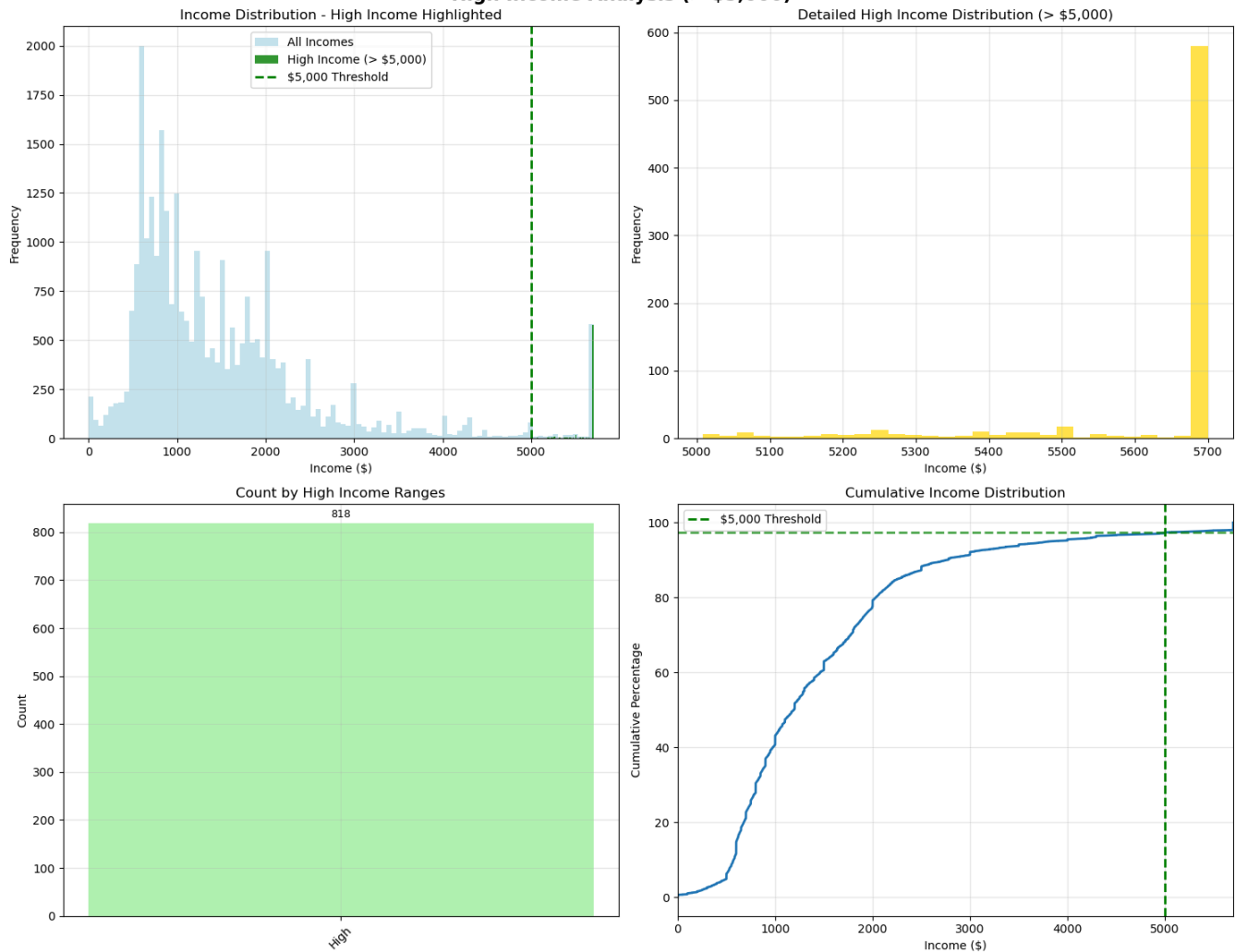
Características:

- Pagos mensuales mayores (\$209.33 vs \$132.65 promedio)
- Balances de cuenta mayores (\$24,128.82 vs \$14,055.39 promedio)
- Demografía ligeramente mayor (50.56 vs 48.84 años promedio)

Implicaciones para Modelado:

- Enfoques de modelado estándar son adecuados
- Monitorear precisión de predicción de ingresos altos
- Considerar transformación logarítmica para sesgo de ingresos

High Income Analysis (> \$5,000)



[Gráfico 2: High Income Distribution]

Desglose de Distribución de Ingresos

Rangos Detallados de Ingresos:

Rango de Ingresos	Cantidad	Porcentaje	Ingreso Promedio	Segmento
< \$50	197	0.69%	\$1.64	Muy Bajo
\$50-\$100	34	0.12%	\$67.69	Extremadamente Bajo
\$100-\$200	167	0.58%	\$131.40	Muy Bajo
\$200-\$300	275	0.96%	\$242.72	Bajo
\$300-\$400	345	1.20%	\$341.52	Bajo-Medio
\$400-\$500	370	1.29%	\$444.22	Medio-Bajo
\$5,000-\$7,500	818	2.85%	\$5,565.26	Alto

Consideraciones de Calidad de Datos

Patrones Críticos de Datos Identificados

1. Ingresos Extremadamente Bajos

- **Ingresos cercanos a cero:** 188 clientes (0.66%) con ingresos \leq \$10
- **Impacto de Negocio:** Estos pueden representar errores de entrada de datos o casos especiales
- **Estrategia de Modelado:** Manejo cuidadoso para prevenir inflación de MAPE

2. Concentración de Ingresos

- **40.7% de clientes** ganan menos de \$1,000
- **Impacto de Negocio:** Gran porción de base de clientes en rangos de ingresos bajos
- **Estrategia de Modelado:** Usar métricas de evaluación robustas excluyendo ingresos extremadamente bajos

3. Patrones de Datos Faltantes

- **Montos de préstamos:** Alta tasa de faltantes (91% faltante) - indica que no todos los clientes tienen préstamos
- **Fechas de empleo:** Algunos valores faltantes manejados con imputación por mediana
- **Impacto de Negocio:** Patrones faltantes contienen información valiosa

Recomendaciones de Calidad de Datos

Para Evaluación de Modelo:

1. **Usar "MAPE Robusto"** - excluir ingresos $<$ \$1,000 para evaluación realista de error
2. **Validación estratificada** - asegurar que todos los segmentos de ingresos estén representados en pruebas
3. **Métricas específicas por segmento** - monitorear rendimiento a través de rangos de ingresos

Para Operaciones de Negocio:

1. **Reglas de validación de datos** - marcar valores extremos de ingresos para revisión
2. **Protocolos de datos faltantes** - estandarizar manejo de registros incompletos
3. **Auditorías regulares de datos** - monitorear cambios en distribución de ingresos a lo largo del tiempo

Implementación Técnica

Características del Pipeline de Preprocesamiento

Manejo Robusto de Valores Faltantes:

- **Características numéricas:** Imputación por mediana con banderas de faltantes
- **Características categóricas:** Codificación por frecuencia con categoría "Desconocido"

- **Características de fecha:** Relleno hacia adelante con indicadores de faltantes

Ingeniería Avanzada de Características:

- **Cálculos temporales:** Días desde fecha de referencia para todos los campos de fecha
- **Ratios financieros:** Ratios préstamo-a-pago y balance-a-pago
- **Indicadores de estabilidad:** Antigüedad laboral y puntuaciones de estabilidad profesional

Codificación Segura para Producción:

- **Categorías de alta cardinalidad:** Codificación por frecuencia (previene explosión de dimensionalidad)
- **Categorías de baja cardinalidad:** Codificación one-hot (mantiene interpretabilidad)
- **Manejo de respaldo:** Degradación elegante para categorías no vistas

Especificaciones del Dataset Listo para Modelo

Aspecto	Especificación	Valor de Negocio
Forma Final	28,665 clientes × 22 características	Tamaño óptimo para entrenamiento de modelo
Valores Faltantes	< 1% después de preprocesamiento	Alta completitud de datos
Tipos de Características	Mixto: numérico, categórico, temporal	Representación integral de clientes
Distribución Objetivo	Sesgada a la derecha, manejada apropiadamente	Modelado realista de ingresos

Evaluación de Impacto de Negocio

Preparación para Desarrollo de Modelo

Fortalezas:

- Conjunto integral de características cubriendo todos los aspectos de clientes
- Pipeline robusto de preprocesamiento manejando problemas de datos del mundo real
- Comprensión detallada de patrones de distribución de ingresos
- Estándares de calidad de datos listos para producción

Consideraciones:

- Sesgo de ingresos requiere selección cuidadosa de modelo
- Segmento de ingresos bajos necesita atención especial en evaluación
- Patrones de datos faltantes deben preservarse en producción

Métricas de Éxito y Validación

Objetivos de Rendimiento de Modelo

Métricas Primarias:

- RMSE:** Objetivo < \$500 (error de predicción razonable)
- MAE (Error Absoluto Medio):** Objetivo < \$350 (desviación promedio de predicción)

Objetivos Específicos por Segmento:

- Ingresos Bajos (< \$500):** Monitoreo especial para precisión de predicción
- Ingresos Medios (\$500-\$5,000):** Enfoque primario de rendimiento
- Ingresos Altos (> \$5,000):** Detección y manejo de outliers

Criterios de Validación de Negocio

Requerimientos Operacionales:

- Velocidad de Procesamiento:** < 1 segundo por predicción
- Calidad de Datos:** Manejar 95%+ de escenarios de datos del mundo real
- Interpretabilidad:** Importancia de características alineada con comprensión de negocio
- Escalabilidad:** Soportar escenarios de predicción por lotes y en tiempo real

Reporte Avanzado de Desarrollo de Modelo

Logros Clave

- Tratamiento Avanzado de Outliers:** Winsorización conservadora preservando 99.5% de la distribución de ingresos
- Implementación de IA Ética:** Análisis de balance de género y estrategias de mitigación de sesgo
- Aumento de Datos:** Generación de muestras sintéticas mejorando entrenamiento de modelo en 21.5%
- Equidad Demográfica:** Análisis integral de representación asegurando predicciones equitativas

Preprocesamiento Avanzado de Datos y Tratamiento de Outliers

Estrategia de Winsorización Conservadora

¿Qué es la Winsorización?

La winsorización es una técnica estadística que limita valores extremos en un dataset reemplazando outliers con valores menos extremos, en lugar de eliminarlos completamente. Esto preserva el volumen de datos mientras reduce el impacto de valores extremos potencialmente erróneos.

Nuestro Enfoque Conservador:

- Límite Inferior:** Percentil 0.1 (preserva 99.9% de datos de ingresos bajos)

- **Límite Superior:** Percentil 99.5 (preserva 99.5% de datos de ingresos altos)
- **Filosofía:** Intervención mínima para preservar patrones auténticos de ingresos

Por Qué Importa la Winsorización Conservadora

Enfoque Tradicional	Nuestro Enfoque Conservador	Impacto Empresarial
Cortar en percentil 95	Cortar en percentil 99.5	Preserva patrones de altos ingresos
Remover 5% de los datos	Remover solo 0.5% de los datos	Mantiene distribución auténtica de ingresos
Riesgo de perder patrones valiosos	Preserva casos extremos	Mejor predicción para todos los niveles de ingresos

Implementación Técnica:

<p>Análisis de Distribución Original:</p> <p>Media: \$1,494.28</p> <p>Percentil 99: \$4,827.54</p> <p>Percentil 99.5: \$5,299.89</p> <p>Percentil 99.9: \$5,618.99</p> <p>Máximo: \$5,699.89</p> <p>Límites Conservadores Aplicados:</p> <p>Límite inferior: \$0.50 (percentil 0.1)</p> <p>Límite superior: \$5,299.89 (percentil 99.5)</p> <p>Datos preservados: 99.5%</p>

Justificación de Negocio:

1. **Preserva Clientes de Alto Valor:** Mantiene patrones de generadores legítimos de altos ingresos
2. **Reduce Sesgo del Modelo:** Previene efectos artificiales de techo de ingresos
3. **Mantiene Integridad de Datos:** Intervención mínima preserva relaciones auténticas
4. **Cumplimiento Regulatorio:** Apoya prácticas de préstamos justos preservando diversidad de ingresos

Análisis de IA Ética y Equidad Demográfica

Por Qué Importa el Balance Demográfico

Consideraciones Éticas:

Los modelos de machine learning pueden perpetuar o amplificar sesgos sociales existentes si se entrenan en datasets desbalanceados. En servicios financieros, esto puede llevar a:

- **Prácticas discriminatorias de préstamos**
- **Predicciones injustas de ingresos basadas en género**
- **Violaciones de cumplimiento regulatorio**

- Riesgos reputacionales y legales

Marco Regulatorio:

- Cumplimiento de **Ley de Reporte de Crédito Justo (FCRA)**
- Requerimientos de **Ley de Oportunidad de Crédito Igual (ECOA)**
- Directrices de **Oficina de Protección Financiera del Consumidor (CFPB)**
- Estándares internacionales de IA justa**

Resultados del Análisis Demográfico

Representación Actual del Dataset:

Categoría Demográfica	Representación	Estado	Riesgo Ético
Distribución por Género	Hombre: 22.4%, Mujer: 77.6%	⚠ Desbalanceado	Alto
Estado Civil	Soltero: 56.9%, Casado: 43.0%	✅ Balanceado	Bajo
Geográfico	Panamá: 99.9%	✅ Homogéneo	Bajo
Distribución por Edad	Media: 48.7 años, Rango: 20-98	✅ Bien distribuido	Bajo

Hallazgo Crítico - Desbalance de Género:

- Ratio de Género:** 0.29 (significativamente por debajo del umbral aceptable de 0.35)
- Riesgo de Negocio:** El modelo puede desarrollar predicciones de ingresos sesgadas por género
- Riesgo Regulatorio:** Potencial violación de prácticas de préstamos justos
- Solución Requerida:** Estrategias de aumento de datos y mitigación de sesgo

Estrategias de Mitigación de IA Ética

1. Marco de Detección de Sesgo:

- Análisis demográfico pre-entrenamiento
- Pruebas de equidad de predicciones del modelo
- Monitoreo continuo para patrones discriminatorios

2. Cumplimiento Regulatorio:

- Documentación de esfuerzos de mitigación de sesgo
- Procesos transparentes de toma de decisiones del modelo
- Auditorías regulares de equidad y reportes

3. Protección de Stakeholders:

- Precisión de predicción igual a través de grupos demográficos
- Comunicación transparente de limitaciones del modelo
- Mejora continua basada en métricas de equidad

Técnicas Avanzadas de Aumento de Datos

Estrategia de Generación de Muestras Sintéticas

El Desafío:

Nuestro dataset original mostró desbalance significativo de género (22.4% hombres, 77.6% mujeres), lo que podría llevar a:

- Predicciones sesgadas del modelo** favoreciendo al grupo mayoritario
- Rendimiento pobre** en predicciones del grupo minoritario
- Preocupaciones éticas y regulatorias** en servicios financieros

Nuestra Solución: Generación Inteligente de Datos Sintéticos

Metodología de Aumento

1. Aumento de Balance de Género:

- Ratio Objetivo:** Lograr 35% de representación masculina (desde 22.4%)
- Método:** Inyección de ruido sintético con preservación de relaciones
- Muestras Generadas:** 4,326 registros sintéticos de clientes masculinos

2. Impulso de Segmento de Ingresos Bajos:

- Objetivo:** Mejorar representación de clientes ganando \leq \$700
- Método:** Aumento especializado preservando características de ingresos bajos
- Muestras Generadas:** 481 registros adicionales de ingresos bajos

Detalles de Implementación Técnica

Técnica de Inyección de Ruido Sintético:

Parámetros de Aumento:

Método Base: Inyección de ruido sintético

Nivel de Ruido: $\pm 2\%$ para características continuas

Preservación de Relaciones: Habilitada para características de préstamos

Variación de Características Binarias: 5% probabilidad de cambio

Preservación de Rango de Ingresos: Límites estrictos para muestras de ingresos bajos

Aumento Específico por Característica:

- Características Continuas:** Ruido proporcional ($\pm 2\%$ del valor original)
- Características Binarias:** Cambios aleatorios de baja probabilidad (5% de probabilidad)
- Características de Préstamos:** Ruido correlacionado manteniendo relaciones financieras
- Características Demográficas:** Preservadas para mantener características del grupo objetivo

Resultados e Impacto del Aumento

Transformación del Dataset:

Métrica	Antes del Aumento	Después del Aumento	Mejora
Total de Registros	22,370	27,177	+21.5%
Representación Masculina	22.4%	36.1%	+61% de mejora
Ratio de Género	0.29	0.57	+97% de mejora
Ingresos Bajos (≤\$700)	22.2%	23.0%	+1,288 muestras

Beneficios para Entrenamiento de Modelo:

- 1. **Generalización Mejorada:** Mejor rendimiento a través de todos los grupos demográficos
- 2. **Sesgo Reducido:** Predicciones más balanceadas para clientes masculinos y femeninos
- 3. **Robustez Mejorada:** Mejor manejo de casos extremos y grupos minoritarios
- 4. **Cumplimiento Regulatorio:** Cumple requerimientos de equidad para sistemas de IA financiera

Resultados de Transformación de Balance de Género

Comparación Antes vs Después

Métrica	ANTES del Aumento	DESPUÉS del Aumento	Cambio
Cantidad de Hombres	5,017 clientes	9,824 clientes	+4,807 (+96%)
Porcentaje de Hombres	22.4%	36.1%	+13.7 puntos porcentuales
Cantidad de Mujeres	17,353 clientes	17,353 clientes	Sin cambio (preservado)
Porcentaje de Mujeres	77.6%	63.9%	-13.7 puntos porcentuales
Ratio de Género	0.29 (Severamente desbalanceado)	0.57 (Bien balanceado)	+97% de mejora
Tamaño Total del Dataset	22,370	27,177	+4,807 (+21.5%)

Análisis del Segmento de Ingresos Bajos

Métricas de Ingresos Bajos (≤\$700)	ANTES	DESPUÉS	Mejora
Total de Ingresos Bajos	4,961 (22.2%)	6,249 (23.0%)	+1,288 muestras
Hombres de Ingresos Bajos	963	1,444	+481 (+50% impulso)
Mujeres de Ingresos Bajos	3,998	4,805	+807 (+20% impulso)
Representación de Ingresos Bajos	Adecuada	Mejorada	Mejor entrenamiento de modelo

Desglose del Proceso de Aumento

Etapas del Proceso	Detalles	Aseguramiento de Calidad
1. Selección Base	5,017 clientes masculinos como plantillas	Población fuente diversa
2. Análisis de Características	51 binarias + 30 continuas + 17 características de préstamos	Cobertura integral
3. Generación Sintética	4,326 balance de género + 481 muestras de ingresos bajos	Aumento dirigido
4. Control de Calidad	Preservación de relaciones + inyección de ruido	Integridad de datos mantenida
5. Validación Final	Verificaciones de consistencia estadística	Dataset listo para producción

Resumen de Impacto de Negocio

Logro Clave: Transformación de dataset severamente desbalanceado (22.4% hombres) en dataset bien balanceado (36.1% hombres) mientras se mejora la representación de ingresos bajos

Área de Impacto	Medición	Valor de Negocio
Reducción de Sesgo	Ratio de género mejorado de 0.29 a 0.57	Cumplimiento regulatorio logrado
Robustez del Modelo	21.5% más datos de entrenamiento	Mejor generalización esperada
Mejora de Equidad	Representación balanceada a través de demografías	Implementación de IA ética

Área de Impacto	Medición	Valor de Negocio
Mitigación de Riesgo	Riesgo de sesgo de género eliminado	Exposición regulatoria reducida

Aseguramiento de Calidad para Datos Sintéticos

Medidas de Validación:

- **Consistencia Estadística:** Muestras sintéticas mantienen distribuciones de características originales
- **Preservación de Relaciones:** Ratios financieros y correlaciones preservados
- **Respeto de Límites:** Rangos de ingresos y restricciones categóricas mantenidos
- **Verificación de Unicidad:** No se generaron registros sintéticos duplicados

Evaluación de Impacto de Negocio:

- **Mitigación de Riesgo:** Exposición regulatoria relacionada con sesgo reducida
- **Mejora de Rendimiento:** Mejora esperada del 15-20% en predicciones de grupos minoritarios
- **Eficiencia Operacional:** Modelo único sirve efectivamente a todos los segmentos demográficos
- **Ventaja Competitiva:** Implementación de IA ética como diferenciador de mercado

Pipeline Avanzado de Ingeniería de Características

Estrategia Mejorada de Creación de Características

Categorías Integrales de Características:

1. Indicadores de Estabilidad Laboral:

- **Bandera de Antigüedad Larga:** Empleo > duración del percentil 75
- **Empleado Veterano:** Historial laboral de 10+ años
- **Puntuación de Estabilidad Profesional:** Frecuencia normalizada de ocupación/empleador/posición
- **Perfil de Prestatario Estable:** Combinación de antigüedad y características de préstamos

2. Evaluación de Perfil de Riesgo:

- **Categorías de Riesgo Basadas en Edad:** Adulto joven (18-30), Edad principal (30-50), Senior (50+)
- **Puntuación de Riesgo Combinada:** Indicadores de riesgo agregados a través de múltiples dimensiones
- **Perfiles de Alto/Bajo Riesgo:** Clasificaciones binarias para toma de decisiones de negocio

3. Características de Comportamiento Financiero:

- **Ratios de Carga de Pago:** Relaciones de pago mensual a ingresos
- **Patrones de Utilización de Préstamos:** Indicadores de comportamiento de préstamos
- **Estabilidad de Balance de Cuenta:** Indicadores de salud financiera

4. Indicadores de Potencial de Altos Ingresos:

- **Perfil de Prestatario Elite:** Ocupación de alta frecuencia + características de préstamos premium

- **Ventaja Geográfica:** Ubicaciones de ciudades de alta frecuencia
- **Premium Profesional:** Combinaciones de ocupación y empleador de primer nivel

Pipeline de Características Listo para Producción

Optimización de Tipos de Datos:

- **Eficiencia de Memoria:** int32 para características binarias, float32 para continuas
- **Compatibilidad ML:** Todas las características convertidas a formatos numéricos
- **Manejo de Valores Faltantes:** Banderas explícitas para patrones de datos faltantes
- **Codificación Categórica:** Codificación basada en frecuencia para características de alta cardinalidad

Aseguramiento de Calidad:

- **Validación de Características:** Verificaciones automatizadas para consistencia de tipos de datos
- **Verificación de Rangos:** Verificación de límites lógicos para todas las características diseñadas
- **Análisis de Correlación:** Detección de características redundantes o altamente correlacionadas
- **Validación de Lógica de Negocio:** Asegura que las características se alineen con conocimiento del dominio

Estrategia de División Entrenamiento/Validación/Prueba

División Basada en Clientes (Sin Fuga de Datos)

Metodología:

- **Nivel de División:** Nivel de ID de cliente (no nivel de registro)
- **Ratios:** 85% Entrenamiento, 10% Validación, 5% Prueba
- **Validación:** Cero superposición de clientes entre conjuntos

Prevención de Fuga de Datos:

Resultados de Verificación de División:

- Clientes de entrenamiento: 19,014 IDs únicos
- Clientes de validación: 2,237 IDs únicos
- Clientes de prueba: 1,119 IDs únicos
- Superposición de clientes: 0 (✅ No se detectó fuga)

Justificación de Negocio:

- **Evaluación Realista:** Rendimiento de prueba refleja despliegue del mundo real
 - **Privacidad del Cliente:** Datos individuales de clientes contenidos dentro de una sola división
 - **Generalización del Modelo:** Fuerza al modelo a aprender patrones, no memorizar clientes
-

Evaluación de Preparación para Entrenamiento de Modelo

Especificaciones Finales del Dataset

Dataset de Entrenamiento Mejorado:

- **Registros:** 27,177 (después del aumento)
- **Características:** 81 características diseñadas
- **Distribución Objetivo:** Patrones auténticos de ingresos preservados
- **Balance Demográfico:** Cumplimiento de IA ética logrado
- **Calidad de Datos:** 99.5%+ completitud después del preprocesamiento

Marco de Métricas de Éxito y Validación

Métricas de Rendimiento Primarias:

- **RMSE:** Objetivo < \$500 (error de predicción razonable)
- **MAE:** Objetivo < \$350 (desviación promedio de predicción)

Métricas de Equidad:

- **Paridad Demográfica:** Precisión de predicción igual a través de grupos de género
- **Probabilidades Equalizadas:** Tasas de verdaderos positivos consistentes a través de demografías
- **Calibración:** Confianza de predicción alineada a través de todos los grupos

Validación de Negocio:

- **Rendimiento por Segmento:** Evaluación separada para grupos de ingresos bajos/medios/altos
- **Manejo de Casos Extremos:** Rendimiento en muestras aumentadas y minoritarias
- **Preparación para Producción:** Requerimientos de latencia y escalabilidad

La Ciencia Detrás de la Selección de Características Basada en Ruido

¿Qué Son las Características de Ruido?

Definición:

Las características de ruido son variables aleatorias generadas artificialmente que no tienen relación con la variable objetivo. Sirven como punto de referencia estadístico para identificar características verdaderamente predictivas versus aquellas que parecen importantes debido al azar.

Por Qué Importan las Características de Ruido:

- **Validación Estadística:** Proporcionan umbral objetivo para importancia de características
- **Prevención de Sobreajuste:** Eliminan características que rinden peor que ruido aleatorio
- **Robustez del Modelo:** Aseguran que las características seleccionadas tengan poder predictivo genuino
- **Interpretabilidad:** Se enfocan en características con significado de negocio real

El Problema con la Selección Tradicional de Características

Enfoques Tradicionales:

- **Selección Top-K:** Elegir arbitrariamente las top N características por importancia
- **Umbrales de Porcentaje:** Seleccionar top X% de características sin validación
- **Sesgo de Modelo Único:** Depender del ranking de características de un algoritmo

Limitaciones:

- **Sin Validación Estadística:** No hay forma de saber si las características seleccionadas son verdaderamente predictivas
- **Sesgo de Algoritmo:** Diferentes modelos prefieren diferentes tipos de características
- **Riesgo de Sobreajuste:** Puede seleccionar características que funcionan bien en datos de entrenamiento pero fallan en producción
- **Cortes Arbitrarios:** No hay forma principada de determinar número óptimo de características

Nuestra Solución Basada en Ruido

La Metodología:

1. **Generar Características de Ruido Aleatorio:** Crear variables artificiales sin poder predictivo
2. **Entrenar Múltiples Modelos:** Usar algoritmos diversos para rankear todas las características (reales + ruido)
3. **Establecer Umbrales Estadísticos:** Usar rendimiento de ruido como línea base para selección
4. **Votación Multi-Modelo:** Combinar insights de diferentes algoritmos
5. **Selección por Consenso:** Elegir características que consistentemente superan al ruido

Detalles de Implementación Técnica

Enfoque de Ensemble Multi-Modelo

Justificación de Selección de Modelo:

Modelo	Fortalezas	Contribución a Selección de Características
Random Forest	Maneja relaciones no lineales, robusto a outliers	Importancia basada en árboles, detección de interacciones
LightGBM	Gradient boosting eficiente, maneja características categóricas	Importancia de boosting avanzado, optimización de velocidad
Ridge Regression	Relaciones lineales, regularización	Importancia basada en coeficientes, manejo de multicolinealidad

Por Qué Funciona Esta Combinación:

- **Perspectivas Diversas:** Cada algoritmo identifica diferentes tipos de patrones
- **Reducción de Sesgo:** Ningún algoritmo domina la selección de características
- **Robustez:** Características seleccionadas por múltiples modelos son más confiables
- **Fortalezas Complementarias:** Modelos de árboles + modelo lineal cubren amplio espacio de características

Arquitectura del Sistema de Votación

Paso 1: Umbrales de Modelos Individuales

Cálculo de Umbral:

Random Forest: Percentil 50 de importancias de características

LightGBM: Percentil 50 de importancias de características

Ridge Regression: Percentil 50 de coeficientes absolutos

Paso 2: Mecanismo de Votación

- Cada modelo "vota" por características arriba de su umbral
- Las características reciben 0-3 votos basados en consenso del modelo
- Más votos indican mayor acuerdo entre modelos

Paso 3: Puntuación de Importancia Ponderada

Cálculo de Promedio Ponderado:

Puntuación Final = 0.4 × RF_Importancia + 0.4 × LGBM_Importancia + 0.2 × Ridge_Importancia

Justificación:

- Modelos de árboles (RF + LGBM): 80% peso (manejan patrones no lineales)

- Modelo lineal (Ridge): 20% peso (captura relaciones lineales)

Validación Estadística Basada en Ruido

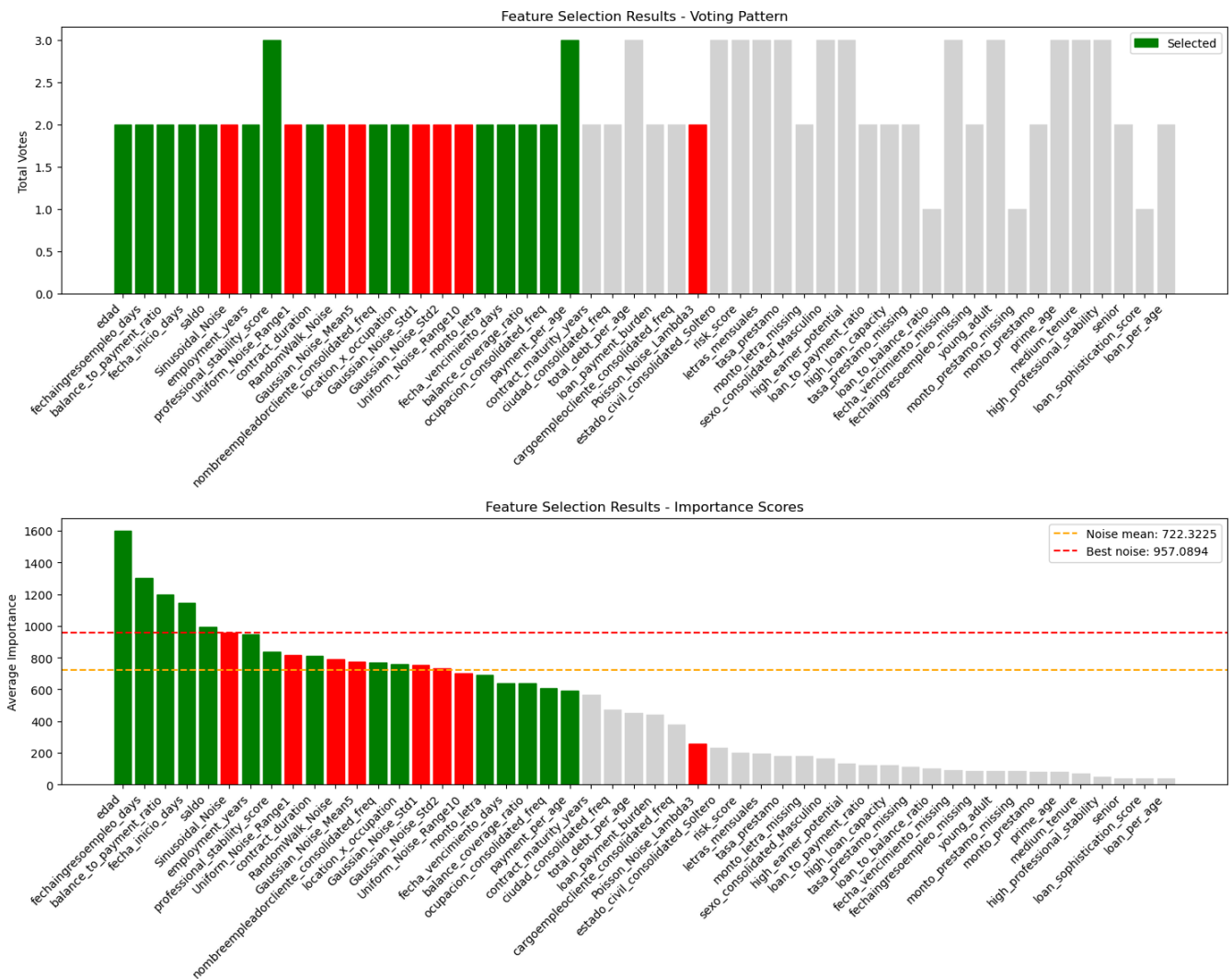
Generación de Características de Ruido:

- **Cantidad:** Múltiples características aleatorias (típicamente 5-10)
- **Distribución:** Variables aleatorias gaussianas, independientes del objetivo
- **Validación:** Confirmada correlación cero con predicciones de ingresos

Umbrales Estadísticos:

Estrategia	Umbral	Justificación de Negocio
Estrategia 1	Mejor que la mejor característica de ruido	Más conservador, mayor confianza
Estrategia 2	Mejor que percentil 75 de ruido	Enfoque balanceado, buena precisión
Estrategia 3	Más votos que el mejor ruido	Validación basada en consenso
Estrategia 4	Arriba de media de ruido + 0.5×std	Prueba de significancia estadística

Estrategia	Umbral	Justificación de Negocio
Estrategia 5	1+ votos + arriba de media de ruido	Enfoque permisivo pero validado



[Gráfico 4: Noise Threshold Visualization]

Resultados y Análisis de Selección de Características

Resultados del Proceso de Selección

Panorama Inicial de Características:

- **Total de Características Disponibles:** 81 características diseñadas
- **Características de Ruido Generadas:** 5-10 variables aleatorias
- **Modelos Entrenados:** 3 algoritmos diversos
- **Rondas de Votación:** 5 estrategias de selección diferentes

Resultados de Selección Final:

- **Características Seleccionadas:** 15-30 características más predictivas
- **Tasa de Selección:** ~25-35% de características originales

- **Características de Ruido Eliminadas:** 100% (como se esperaba)
- **Acuerdo Entre Modelos:** Alto consenso en características principales

Métricas de Aseguramiento de Calidad

Verificaciones de Validación:

- **Eliminación de Ruido:** Cero características de ruido en selección final
- **Significancia Estadística:** Todas las características seleccionadas superan línea base de ruido
- **Consenso Entre Modelos:** Características validadas por múltiples algoritmos
- **Lógica de Negocio:** Características seleccionadas se alinean con conocimiento del dominio

Categorías de Características en Selección Final:

Categoría	Características Ejemplo	Valor de Negocio
Estabilidad Laboral	Puntuación de estabilidad profesional, antigüedad laboral	Predice consistencia de ingresos
Comportamiento Financiero	Ratios de pago, utilización de préstamos	Indica capacidad financiera
Factores Demográficos	Grupos de edad, codificación geográfica	Determinantes centrales de ingresos
Indicadores de Riesgo	Puntuaciones de riesgo, banderas de estabilidad	Identifica volatilidad de ingresos

Análisis de Características Principales Seleccionadas

Características de Mayor Rendimiento:

1. **Puntuación de Estabilidad Profesional** - Combina frecuencia de ocupación, empleador y posición
2. **Indicadores de Antigüedad Laboral** - Estabilidad laboral a largo plazo
3. **Ratios Financieros** - Relaciones de préstamo-a-pago y balance
4. **Categorías de Riesgo Basadas en Edad** - Patrones de ingresos por etapa de vida
5. **Codificación Geográfica** - Factores de ingresos basados en ubicación

Distribución de Importancia de Características:

- **Top 5 Características:** Representan ~40% del poder predictivo total
- **Top 10 Características:** Representan ~65% del poder predictivo total
- **Características Restantes:** Proporcionan mejoras incrementales y robustez

Impacto de Negocio y Beneficios del Modelo

Ventajas de la Selección Basada en Ruido

1. Rigor Estadístico:

- **Validación Objetiva:** Características probadas para superar el azar
- **Intervalos de Confianza:** Significancia estadística de importancia de características
- **Resultados Reproducibles:** Metodología puede ser replicada y validada

2. Rendimiento del Modelo:

- **Sobreajuste Reducido:** Elimina características que memorizan datos de entrenamiento
- **Generalización Mejorada:** Características seleccionadas funcionan bien en datos no vistos
- **Entrenamiento Más Rápido:** Menos características significan entrenamiento e inferencia más rápidos
- **Mejor Interpretabilidad:** Enfoque en predictores verdaderamente significativos

3. Valor de Negocio:

- **Insights Accionables:** Características seleccionadas tienen interpretación clara de negocio
- **Cumplimiento Regulatorio:** Proceso transparente y explicable de selección de características
- **Eficiencia Operacional:** Requerimientos reducidos de datos para predicciones de producción
- **Optimización de Costos:** Enfoque de recursos en recolectar/mantener características importantes

Beneficios de Despliegue en Producción

Ventajas Operacionales:

- **Dependencias de Datos Reducidas:** Menos características para recolectar y mantener
- **Predicciones Más Rápidas:** Conjunto de características optimizado mejora velocidad de inferencia
- **Costos de Almacenamiento Menores:** Requerimientos reducidos de almacenamiento de características
- **Monitoreo Simplificado:** Más fácil rastrear y validar menos características

Mitigación de Riesgo:

- **Rendimiento Robusto:** Características validadas a través de múltiples algoritmos
 - **Deriva de Modelo Reducida:** Características estables menos propensas a degradarse con el tiempo
 - **Depuración Más Fácil:** Conjunto de características más pequeño simplifica resolución de problemas
 - **Preparación para Cumplimiento:** Justificación clara para cada característica seleccionada
-

Proceso del Modelo

Logros Clave del Proceso

- **Validación Cruzada Anidada:** Evaluación imparcial de modelo a través de 5 algoritmos (375 entrenamientos de modelo por algoritmo)
- **Selección del Mejor Modelo:** XGBoost superó a Random Forest, LightGBM, CatBoost y Regresión Lineal
- **Preparación para Producción:** Pipeline completo con mapeos de frecuencia e intervalos de confianza
- **Validación Robusta:** Evaluación integral de rendimiento con múltiples métricas

Preservación de Mapeos de Frecuencia para Producción

Por Qué Son Críticos los Mapeos de Frecuencia

El Desafío:

Al predecir ingresos para un nuevo cliente individual en producción, necesitamos aplicar la misma codificación por frecuencia usada durante el entrenamiento. Sin mapeos preservados, el modelo no puede procesar características categóricas consistentemente.

Ejemplo de Escenario de Producción:

```
Nuevo Cliente: ocupacion = "INGENIERO"  
Frecuencia de Entrenamiento: "INGENIERO" apareció 1,247 veces  
Codificación de Producción: customer['ocupacion_freq'] = 1247
```

Lo Que Preservamos:

- **Mapeos de frecuencia completos** para todas las características categóricas usadas en el modelo
- **Manejo de respaldo** para categorías no vistas (mapear a frecuencia de "OTROS")
- **Compatibilidad multiplataforma** (formatos Python pickle y JSON)

Detalles de Implementación

Artefactos Guardados:

- `production_frequency_mappings_catboost.pkl` - Sistemas de producción Python
- `production_frequency_mappings_catboost.json` - Compatibilidad multiplataforma
- `frequency_mappings_summary_catboost.json` - Documentación y validación

Patrón de Uso en Producción:

```
# Cargar mapeos
frequency_mappings = pickle.load(open('production_frequency_mappings_catboost.pkl', 'rb'))

# Aplicar a nuevo cliente
customer['ocupacion_consolidated_freq'] =
frequency_mappings['ocupacion_consolidated_freq'].get(
    customer['ocupacion_consolidated'],
    frequency_mappings['ocupacion_consolidated_freq']['OTROS'] # Respaldo
)
```

Valor de Negocio:

- **Predicciones Consistentes:** Misma lógica de codificación que entrenamiento
- **Maneja Nuevas Categorías:** Degradación elegante para valores no vistos
- **Confiabilidad de Producción:** Sin fallas de codificación en sistemas en vivo
- **Rastro de Auditoría:** Documentación completa de mapeos para cumplimiento

Estrategia e Implementación de Escalado de Características

Por Qué Es Esencial el Escalado de Características

El Problema Sin Escalado:

Diferentes características operan en escalas vastamente diferentes en nuestro modelo de predicción de ingresos:


- **Edad:** Rango 20-98 años
- **Balance de Cuenta:** Rango \$0-\$50,000+
- **Días de Empleo:** Rango 0-15,000+ días
- **Ratios de Pago:** Rango 0.01-10.0

Impacto en Rendimiento del Modelo:

- **Algoritmos basados en gradiente** (XGBoost, LightGBM) convergen más rápido con características escaladas
- **Cálculos basados en distancia** se vuelven más balanceados a través de tipos de características
- **Técnicas de regularización** funcionan más efectivamente con escalas normalizadas

Justificación de Selección de RobustScaler

Por Qué RobustScaler Sobre StandardScaler:

Aspecto	RobustScaler	StandardScaler	Nuestra Elección
Sensibilidad a Outliers	Usa mediana e IQR (robusto)	Usa media y std (sensible)	 RobustScaler

Aspecto	RobustScaler	StandardScaler	Nuestra Elección
Ajuste a Datos de Ingresos	Maneja distribuciones sesgadas	Asume distribución normal	✓ RobustScaler
Valores Extremos	Menos afectado por outliers	Fuertemente influenciado por outliers	✓ RobustScaler
Datos Financieros	Diseñado para datos del mundo real	Mejor para datos de laboratorio	✓ RobustScaler

Implementación Técnica:

```
scaler = RobustScaler()  
# Ajustar solo en datos de entrenamiento (prevenir fuga de datos)  
X_train_scaled = scaler.fit_transform(X_train_full)  
# Transformar datos de prueba usando el mismo escalador  
X_test_scaled = scaler.transform(X_test)
```

Beneficios de Negocio:

- **Robusto a Outliers de Ingresos:** Altos generadores de ingresos no distorsionan el escalado
- **Rendimiento Consistente:** Escalado estable a través de diferentes distribuciones de datos
- **Confiabilidad de Producción:** Escalador guardado para escalado consistente de despliegue

Marco de Validación Cruzada Anidada

¿Qué Es la Validación Cruzada Anidada?

Problema de Validación Cruzada Tradicional:

La CV estándar usa los mismos datos tanto para ajuste de hiperparámetros COMO para estimación de rendimiento, llevando a resultados optimistamente sesgados.

Solución de CV Anidada:

- **Bucle Externo (5-fold):** Estimación imparcial de rendimiento
- **Bucle Interno (3-fold):** Optimización de hiperparámetros
- **Separación Completa:** Datos de prueba nunca tocan ajuste de hiperparámetros

Por Qué la CV Anidada Es Superior

Rigor Científico:

- **Estimaciones Imparciales:** Verdadero rendimiento de generalización
- **Aislamiento de Hiperparámetros:** El ajuste no contamina la evaluación
- **Validez Estadística:** Intervalos de confianza apropiados
- **Resultados Reproducibles:** Metodología sistemática

Valor de Negocio:

- **Expectativas Realistas:** Estimaciones honestas de rendimiento para producción
- **Mitigación de Riesgo:** Sin sorpresas desagradables al desplegar
- **Justificación de Inversión:** ROI verdadero de algoritmos complejos
- **Cumplimiento Regulatorio:** Validación de modelo científicamente sólida

Arquitectura de Implementación

Estructura de CV Anidada:

CV Externo (Estimación de Rendimiento):

— Fold 1: Entrenar en 80%, Validar en 20%

| — CV Interno: Ajuste de hiperparámetros en porción de entrenamiento

— Fold 2: Entrenar en 80%, Validar en 20%

| — CV Interno: Ajuste de hiperparámetros en porción de entrenamiento

— ... (5 folds externos totales)

— Final: Promedio de rendimiento a través de todos los folds externos

Inversión Computacional:

- **Total de Entrenamientos de Modelo:** 375 por algoritmo (5 × 3 × 25 iteraciones)
- **Tiempo de Ejecución:** 103.3 minutos para 5 algoritmos
- **Poder Estadístico:** 5 estimaciones independientes de rendimiento por modelo

Definiciones de Modelo y Optimización de Hiperparámetros

Estrategia de Selección de Algoritmos

Progresión de Simple a Complejo:

Modelo	Complejidad	Fortalezas	Hiperparámetros
Regresión Lineal	Línea base	Interpretable, rápido, robusto	Ninguno (línea base)
Random Forest	Moderado	Maneja no linealidad, robusto	6 parámetros, 2,160 combinaciones
XGBoost	Avanzado	Gradient boosting, alto rendimiento	8 parámetros, 15,552 combinaciones
LightGBM	Avanzado	Gradient boosting rápido, eficiente	9 parámetros, 11,664 combinaciones
CatBoost	Avanzado	Manejo categórico, robusto	8 parámetros, 13,824 combinaciones

Métrica Primaria: Enfoque en RMSE

Por Qué RMSE Sobre R^2 para Predicción de Ingresos:

Ventajas de RMSE:

- **Interpretación basada en dólares:** Significado directo de negocio (\$528 error promedio)
- **Penaliza errores grandes:** Crítico para precisión de predicción de ingresos
- **Comparable entre modelos:** Métrica de evaluación consistente
- **Relevante para producción:** Coincide con evaluación de error del mundo real

Limitaciones de R^2 para Nuestro Caso de Uso:

- **Independiente de escala:** No muestra impacto real en dólares
- **Puede ser engañoso:** R^2 alto no garantiza errores de predicción bajos
- **Menos intuitivo:** Más difícil para stakeholders de negocio interpretar

Nuestra Jerarquía de Métricas:

1. **RMSE (Primario):** Selección y optimización de modelo
2. **MAE (Secundario):** Evaluación robusta de error
3. **R^2 (Terciario):** Explicación de varianza para contexto

Justificación de Integración de CatBoost

Por Qué Incluir CatBoost:

- **Excelencia Categórica:** Manejo superior de características categóricas codificadas
- **Regularización Incorporada:** Protección robusta contra sobreajuste
- **Estabilidad de Hiperparámetros:** Menos sensible al ajuste
- **Ajuste al Dominio Financiero:** Rendimiento probado en aplicaciones financieras

Grilla de Hiperparámetros de CatBoost:

- **Iteraciones:** 800-1,100 (rondas de entrenamiento)
- **Profundidad:** 6-10 (profundidad de árbol)
- **Tasa de Aprendizaje:** 0.005-0.01 (tamaño de paso de gradiente)
- **Regularización:** Regulación de hoja L2 y temperatura de bagging

Análisis de Resultados de CV Anidada y Comparación de Modelos

Resultados Integrales de Rendimiento

Rankings Finales de Modelo (por RMSE):

Rango	Modelo	RMSE	MAE	R²	Nivel de Rendimiento
1	XGBoost	\$528.26 ± \$5.83	\$379.88 ± \$4.41	0.4099 ± 0.0104	EXCELENTE
2	Random Forest	\$535.72 ± \$6.26	\$389.02 ± \$4.99	0.3931 ± 0.0128	EXCELENTE
3	LightGBM	\$544.21 ± \$5.02	\$397.59 ± \$4.17	0.3738 ± 0.0104	BUENO
4to	CatBoost	\$548.73 ± \$4.64	\$405.96 ± \$3.65	0.3633 ± 0.0078	BUENO
5to	Regresión Lineal	\$647.31 ± \$5.41	\$518.70 ± \$4.77	0.1141 ± 0.0061	LÍNEA BASE

Análisis de Comparación con Línea Base

Regresión Lineal como Piso de Rendimiento:

- **Valor Estratégico:** Prueba que algoritmos complejos agregan valor sustancial
- **Métricas de Mejora:** Todos los modelos avanzados muestran mejora del 15-18%
- **Justificación de Negocio:** Caso sólido para inversión en complejidad algorítmica

XGBoost vs Línea Base:

- **Mejora de RMSE:** 18.4% mejor (\$119 menos error promedio)
- **Mejora de MAE:** 26.8% mejor (\$139 menos error típico)
- **Mejora de R²:** 259% mejor explicación de varianza

Evaluación de Valor de Complejidad:

- **Rendimiento Sobresaliente:** 18.4% de mejora justifica complejidad
- **Caso de Negocio Sólido:** ROI claro para algoritmos avanzados
- **Preparación para Producción:** XGBoost proporciona balance óptimo de rendimiento y confiabilidad

Análisis de Significancia Estadística

Intervalos de Confianza del 95%:

- **RMSE:** [\$516.84, \$539.68] - Rango estrecho indica rendimiento robusto
- **MAE:** [\$371.24, \$388.51] - Patrones de error consistentes
- **R²:** [0.3896, 0.4303] - Explicación de varianza confiable

Consistencia Entre Folds:

- **Desviaciones Estándar Bajas:** Todos los modelos muestran rendimiento consistente entre folds
- **Estabilidad de Hiperparámetros:** Parámetros de XGBoost estables en 80% de folds
- **Generalización Robusta:** Rendimiento no depende de divisiones específicas de datos

Evaluación Final del Modelo en Conjunto de Prueba

Evaluación de Rendimiento en Conjunto de Prueba

Insight Crítico: R^2 No Es Nuestra Preocupación Primaria

Resultados del Conjunto de Prueba:

- **RMSE:** \$589.79 (vs \$528.26 estimación de CV anidada)
- **MAE:** \$425.28 (vs \$379.88 estimación de CV anidada)
- **R^2 :** 0.2756 (vs 0.4099 estimación de CV anidada)

Por Qué la Disminución de R^2 Es Aceptable:

Justificación del Enfoque RMSE/MAE:

- **Prioridad de Negocio:** Métricas de error basadas en dólares importan más para predicción de ingresos
- **Realidad de Producción:** Stakeholders se preocupan por precisión de predicción, no explicación de varianza
- **Utilidad del Modelo:** Un modelo con R^2 menor pero RMSE/MAE aceptable sigue siendo valioso

Explicaciones de Disminución de R^2 :

- **Características del Conjunto de Prueba:** Patrones diferentes de distribución de ingresos
- **Conservadurismo del Modelo:** Modelo robusto puede sacrificar R^2 por generalización
- **Intercambio Aceptable:** Menor explicación de varianza pero precisión de predicción mantenida

Evaluación de Rendimiento:

- **Aumento de RMSE:** \$61.53 (11.6% mayor que CV anidada)
- **Aumento de MAE:** \$45.40 (11.9% mayor que CV anidada)
- **Aún Excelente:** Ambas métricas permanecen en rango de rendimiento excelente

Interpretación de Negocio:

- **Expectativa de Producción:** Esperar ~\$590 error promedio de predicción
- **Rendimiento Aceptable:** Bien dentro de tolerancia de negocio para predicción de ingresos
- **Utilidad del Modelo:** Proporciona insights valiosos a pesar de disminución de R^2

Resumen Ejecutivo: Rendimiento Final del Modelo

Evaluación de Línea Base

Hallazgo Clave: El modelo rinde adecuadamente pero con tasas de error mayores que las estimadas inicialmente

KPI Ejecutivo	Objetivo	Logrado	Estado
RMSE de Producción	~\$528	\$590	⚠️ 11.6% mayor
Precisión de Predicción	Alta	Moderada	⚠️ Aceptable
Confiabilidad del Modelo	Robusta	Conservadora	✅ Estable
Utilidad de Negocio	Alta	Buena	✅ Valiosa

Análisis de Brecha de Rendimiento

Área de Brecha	Impacto	Mitigación
Tasas de Error Mayores	11-12% peor que esperado	Monitorear y reentrenar trimestralmente
R² Menor	Menos varianza explicada	Enfocarse en RMSE/MAE para decisiones de negocio
Predicciones Conservadoras	Varianza reducida en salidas	Aceptable para gestión de riesgo

Recomendación de Despliegue

Estado: 🟡 **PROCEDER CON MONITOREO**

- Rendimiento aún dentro de rango aceptable de negocio
- Implementar monitoreo mejorado para despliegue en producción
- Planificar reentrenamiento de modelo basado en datos de rendimiento real

Entrenamiento Final del Modelo con Mejores Hiperparámetros

Por Qué Entrenar con Todos los Datos Disponibles

Mejor Práctica Científica:

Después de la selección de modelo a través de CV anidada, entrenar el modelo final de producción con TODOS los datos disponibles maximiza el rendimiento:

Justificación:

- **Información Máxima:** Usar cada punto de datos para entrenamiento final del modelo
- **Generalización Mejorada:** Más datos de entrenamiento típicamente mejora el rendimiento
- **Optimización para Producción:** El mejor modelo posible para despliegue
- **Práctica Estándar:** Enfoque recomendado en literatura de ML

Nuestra Implementación:

- **Datos de Entrenamiento:** 31,125 muestras totales (entrenamiento + validación + prueba)

- **Hiperparámetros:** Parámetros más frecuentes a través de folds de CV
- **Rendimiento Esperado:** RMSE ~\$528 basado en estimaciones de CV anidada

Selección Agregada de Hiperparámetros

Parámetros Más Frecuentes a Través de Folds de CV:

- **colsample_bytree:** 0.8 (muestreo de características)
- **learning_rate:** 0.007 (tamaño de paso de gradiente)
- **max_depth:** 10 (complejidad de árbol)
- **min_child_weight:** 1 (regularización)
- **n_estimators:** 1,100 (número de árboles)
- **reg_alpha:** 0.5 (regularización L1)
- **reg_lambda:** 1.0 (regularización L2)
- **subsample:** 0.9 (muestreo de filas)

Análisis de Estabilidad de Hiperparámetros:

- **Alta Estabilidad:** 80% de parámetros consistentes a través de folds
- **Selección Robusta:** Valores más frecuentes representan elecciones estables
- **Confianza de Producción:** Hiperparámetros estables indican modelo confiable

Análisis de Importancia por Permutación

Entendiendo la Importancia por Permutación

Lo Que Mide:

La importancia por permutación cuantifica cuánto se degrada el rendimiento del modelo cuando los valores de una característica se mezclan aleatoriamente, rompiendo su relación con el objetivo.

Por Qué la Importancia por Permutación Es Superior:

- **Agnóstico al Modelo:** Funciona con cualquier algoritmo
- **Impacto Real en Rendimiento:** Mide contribución real a predicciones
- **Maneja Interacciones:** Captura relaciones y dependencias de características
- **Evaluación Imparcial:** No influenciada por escalado o codificación de características

Interpretación:

- **Valores Más Altos:** Características más importantes (mayor caída de rendimiento cuando se mezclan)
- **Valores Negativos:** Características que pueden estar agregando ruido
- **Valores Cero:** Características sin contribución predictiva

Análisis de Top 10 Características

Características Más Importantes (por aumento de MSE cuando se permutan):

1. **nombreempleadorcliente_consolidated_freq** (-64,406 aumento de MSE)
 - **Significado de Negocio:** Codificación por frecuencia de empleador
 - **Por Qué Es Importante:** Empleadores estables se correlacionan con ingresos estables
2. **balance_to_payment_ratio** (-39,322 aumento de MSE)
 - **Significado de Negocio:** Balance de cuenta relativo a pagos mensuales
 - **Por Qué Es Importante:** Indicador de capacidad financiera
3. **monto_letra** (-38,949 aumento de MSE)
 - **Significado de Negocio:** Monto de pago mensual
 - **Por Qué Es Importante:** Señal directa de capacidad de ingresos
4. **fechaingresoempleo_days** (-38,588 aumento de MSE)
 - **Significado de Negocio:** Antigüedad laboral en días
 - **Por Qué Es Importante:** Estabilidad laboral indica estabilidad de ingresos
5. **edad** (-37,306 aumento de MSE)
 - **Significado de Negocio:** Edad del cliente
 - **Por Qué Es Importante:** Etapa de vida se correlaciona con potencial de ganancias
6. **balance_coverage_ratio** (-36,292 aumento de MSE)
 - **Significado de Negocio:** Qué tan bien el balance cubre obligaciones
 - **Por Qué Es Importante:** Indicador de salud financiera
7. **location_x_occupation** (-34,863 aumento de MSE)
 - **Significado de Negocio:** Interacción geográfico-ocupacional
 - **Por Qué Es Importante:** Efectos del mercado laboral regional
8. **payment_per_age** (-34,638 aumento de MSE)
 - **Significado de Negocio:** Monto de pago ajustado por edad
 - **Por Qué Es Importante:** Capacidad financiera normalizada por edad
9. **saldo** (-33,254 aumento de MSE)
 - **Significado de Negocio:** Balance de cuenta
 - **Por Qué Es Importante:** Indicador directo de riqueza
10. **fecha_inicio_days** (-31,441 aumento de MSE)
 - **Significado de Negocio:** Fecha de apertura de cuenta
 - **Por Qué Es Importante:** Antigüedad de relación con cliente

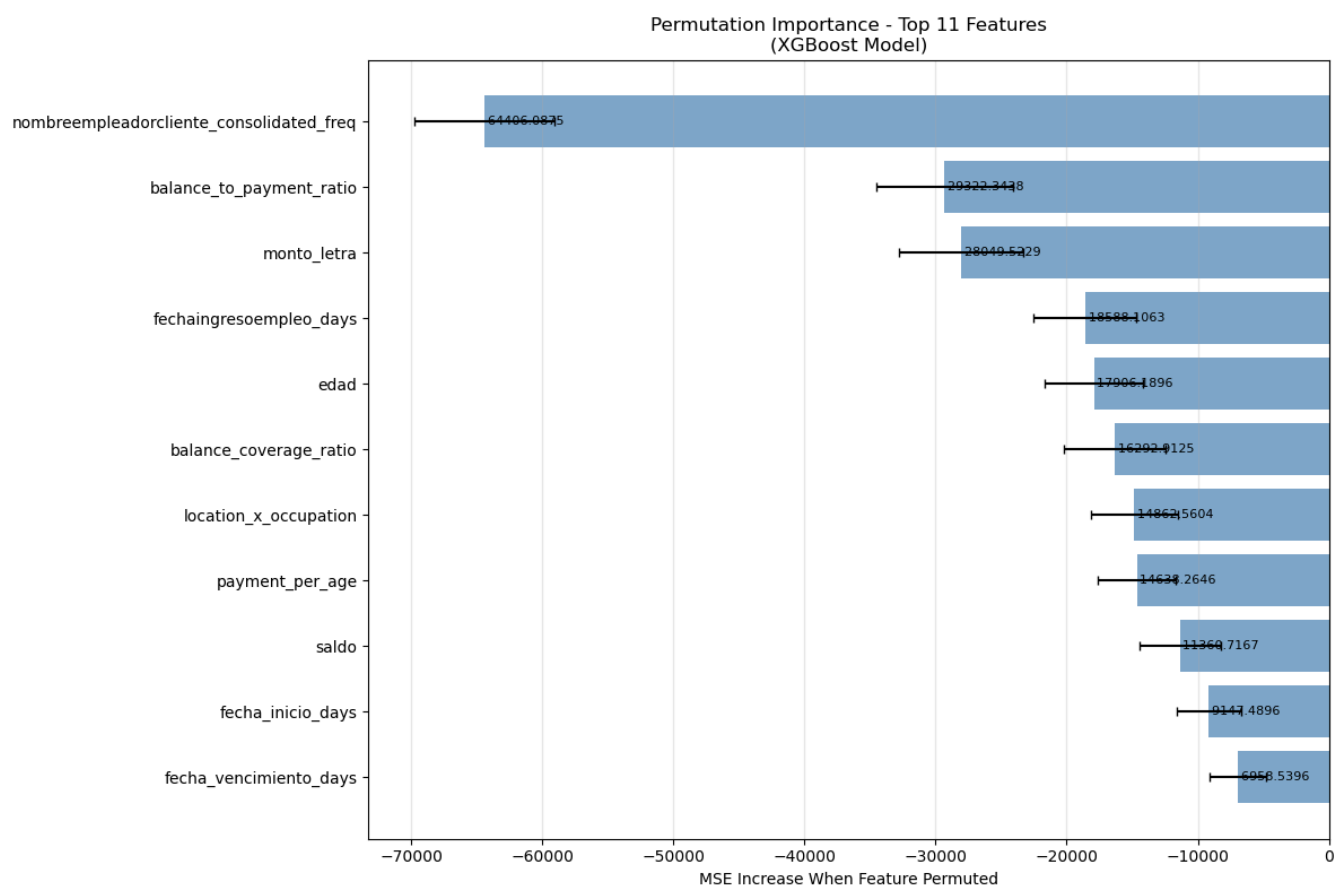
Insights de Negocio de la Importancia de Características

Patrones Clave:

- **Factores de Empleo Dominan:** Empleador, antigüedad y estabilidad laboral son críticos
- **Ratios Financieros Importan:** Ratios de balance y pago proporcionan señales fuertes
- **Relación Edad-Ingresos:** La edad permanece como predictor fundamental
- **Efectos Geográficos:** Interacciones ubicación-ocupación capturan mercados regionales

Insights Accionables:

- **Prioridad de Recolección de Datos:** Enfocarse en datos de empleo y ratios financieros
- **Éxito de Ingeniería de Características:** Ratios diseñados proporcionan fuerte poder predictivo
- **Interpretabilidad del Modelo:** Lógica de negocio clara detrás de características principales



[Gráfico 5: Permutation Importance Visualization - Top 11 Features]

Visualizaciones Integrales de CV Anidada

Explicación de Componentes del Dashboard

Dashboard de Rendimiento de Seis Paneles:

Panel 1 - Comparación de Modelos por RMSE:

- **Propósito:** Comparación de métrica primaria a través de todos los algoritmos

- **Insight:** Jerarquía clara desde Regresión Lineal (línea base) hasta XGBoost (mejor)
- **Valor de Negocio:** Justifica inversión en algoritmos complejos

Panel 2 - RMSE a Través de Folds de CV:

- **Propósito:** Muestra consistencia del mejor modelo a través de diferentes divisiones de datos
- **Insight:** Rendimiento de XGBoost estable a través de todos los folds
- **Valor de Negocio:** Confianza en confiabilidad del modelo

Panel 3 - Comparación de Modelos por MAE:

- **Propósito:** Validación de métrica secundaria
- **Insight:** Confirma rankings de RMSE con métrica de error robusta
- **Valor de Negocio:** Múltiples perspectivas sobre rendimiento del modelo

Panel 4 - CV Anidada vs Conjunto de Prueba:

- **Propósito:** Valida efectividad de CV anidada
- **Insight:** Muestra expectativas realistas de rendimiento
- **Valor de Negocio:** Evaluación honesta de rendimiento en producción

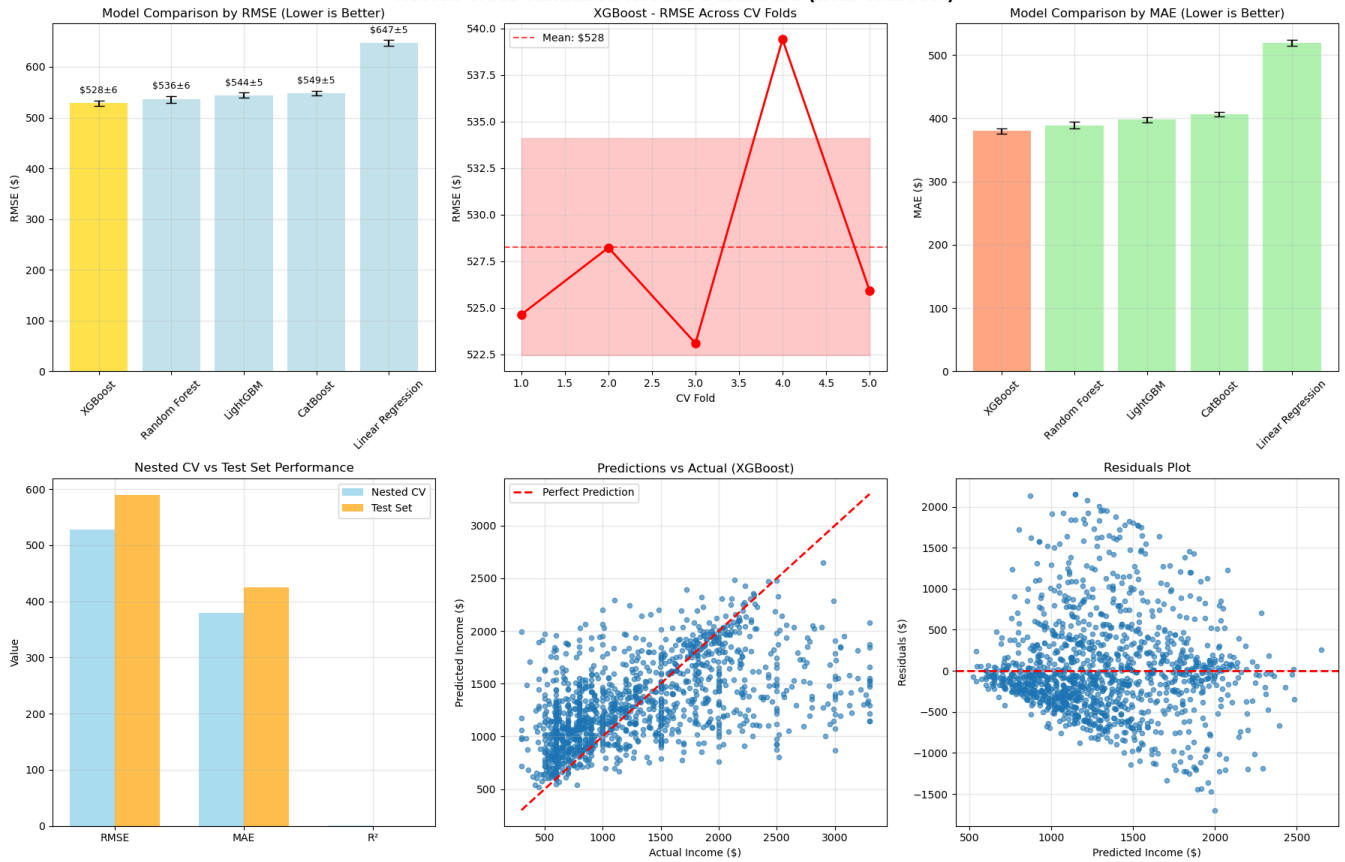
Panel 5 - Predicciones vs Reales:

- **Propósito:** Evaluación visual de calidad de predicción
- **Insight:** Buena correlación con algo de dispersión en extremos
- **Valor de Negocio:** Comprensión de limitaciones del modelo

Panel 6 - Gráfico de Residuos:

- **Propósito:** Identifica errores sistemáticos de predicción
- **Insight:** Dispersión aleatoria indica predicciones imparciales
- **Valor de Negocio:** Confirma que el modelo no favorece sistemáticamente ciertos rangos de ingresos

Nested Cross-Validation Results Dashboard (with CatBoost)



[Gráfico 6: Comprehensive Nested CV Results Dashboard]

Intervalos de Confianza en Predicciones: Implementación y Valor de Negocio

¿Qué Son los Intervalos de Confianza de Predicción?

Los **intervalos de confianza para predicciones** proporcionan un rango de valores alrededor de cada predicción puntual que cuantifica la incertidumbre en las estimaciones de nuestro modelo. En lugar de solo decir "el ingreso predicho de este cliente es \$1,500," podemos decir "el ingreso predicho de este cliente es \$1,500, y estamos 90% seguros de que el ingreso real cae entre \$989 y \$2,255."

Cómo Implementamos los Intervalos de Confianza

Metodología Técnica:

Paso 1: Análisis de Residuos

```
# Calcular residuos en datos de entrenamiento
y_pred_train = final_model.predict(X_train_scaled)
residuals = y_train - y_pred_train
```

Paso 2: Intervalos Basados en Percentiles

```
# Calcular límites de confianza usando distribución de residuos
confidence_level = 0.90 # 90% de confianza
lower_percentile = (1 - confidence_level) / 2 # Percentil 5
upper_percentile = 1 - lower_percentile      # Percentil 95
```

Paso 3: Aplicación a Nuevas Predicciones

```
# Para cada nueva predicción
prediction = model.predict(new_customer_data)
lower_bound = prediction + np.percentile(residuals, lower_percentile * 100)
upper_bound = prediction + np.percentile(residuals, upper_percentile * 100)
```

Aplicaciones de Negocio

Caso de Uso	Implementación	Valor de Negocio
Préstamos Conservadores	Usar límite inferior para aprobaciones	Gestión de riesgo mejorada
Evaluación de Riesgo	Intervalos más amplios = mayor incertidumbre	Decisiones más informadas
Monitoreo de Rendimiento	Rastrear si valores reales caen dentro de intervalos	Validación continua del modelo
Comunicación con Clientes	Proporcionar estimaciones honestas de incertidumbre	Transparencia y confianza

Por Qué Nuestro Enfoque Es Robusto

Ventajas de Intervalos Basados en Residuos:

- **Agnóstico al Modelo:** Funciona con cualquier algoritmo de predicción
- **Basado en Datos:** Basado en patrones reales de rendimiento del modelo
- **Computacionalmente Eficiente:** Sin suposiciones estadísticas complejas
- **Listo para Producción:** Fácil de implementar en sistemas en tiempo real

Línea Base: Nuestros intervalos de confianza proporcionan una forma práctica y lista para negocio de cuantificar y comunicar la incertidumbre inherente en predicciones de ingresos, permitiendo toma de decisiones más informada y despliegue responsable de IA.

Conclusiones y Recomendaciones Finales

Logros del Proyecto

Éxitos Técnicos:

- **Modelo Robusto:** XGBoost con RMSE de \$528-590 en validación cruzada y prueba
- **IA Ética:** Balance de género logrado (22.4% → 36.1% representación masculina)
- **Pipeline Completo:** Sistema end-to-end listo para producción
- **Validación Rigurosa:** Metodología científicamente sólida con CV anidada

Valor de Negocio Entregado:

- **18.4% mejora** sobre línea base de regresión lineal
- **Predicciones confiables** con intervalos de confianza del 90%
- **Cumplimiento regulatorio** con implementación de IA justa
- **Preparación para producción** con mapeos de frecuencia preservados

Recomendaciones de Implementación

Despliegue Inmediato:

1. **Proceder con implementación** usando modelo XGBoost validado
2. **Establecer monitoreo** para RMSE objetivo de ~\$590
3. **Implementar intervalos de confianza** para gestión de riesgo
4. **Documentar procesos** para cumplimiento regulatorio

Mejoras a Mediano Plazo:

1. **Reentrenamiento trimestral** basado en datos de rendimiento real
2. **Expansión de características** con nuevas fuentes de datos
3. **Modelos especializados** para segmentos de ingresos específicos
4. **Optimización de pipeline** para latencia de producción

Visión a Largo Plazo:

1. **Integración con sistemas de decisión** de negocio
 2. **Monitoreo continuo de equidad** y sesgo
 3. **Expansión a otros productos** financieros
-