Títulos

Data Driven Finance: Detectando Patrones en Reclamaciones Bancarias con IA

FinTech Insights: Inteligencia Artificial aplicada a Reclamaciones y Riesgos

Autores:

Luis Carlos Zárate Orjuela

Cristian Eduardo Osorio

Cristian Camilo Duran

Héctor Fabio Londoño Arbeláez

Profesores:

Natalia Betancur

Frank Beltrán

Andrés Sánchez

Fecha: 22/05/2025

Logo institucional:

Breve descripción del problema:

Actualmente, las entidades financieras enfrentan un volumen creciente y diverso de reclamaciones por parte de los usuarios, que ingresan tanto a través de la Superintendencia Financiera como directamente por los canales de atención de los bancos. Estas reclamaciones abarcan múltiples tipologías de casos (errores operativos, cobros indebidos, demoras, entre otros), y su análisis manual resulta ineficiente y poco oportuno.

A pesar de contar con grandes volúmenes de datos, no se aprovechan herramientas analíticas avanzadas para identificar patrones estacionales, tendencias por tipo de entidad, o picos asociados a eventos específicos (como cambios regulatorios o campañas comerciales). Esta falta de análisis profundo impide anticipar problemáticas recurrentes, diseñar respuestas proactivas, y mejorar la experiencia del cliente.

Objetivo general y especifico

Objetivo General:

Desarrollar un modelo de análisis basado en inteligencia artificial que permita identificar patrones, estacionalidades y afectaciones en las reclamaciones presentadas por los usuarios a entidades financieras, con el fin de generar información útil para la toma de decisiones y la mejora del servicio al cliente.

Objetivos Específicos:

1. Integrar y depurar la base de datos de reclamaciones provenientes de la Superintendencia Financiera y de las entidades bancarias, asegurando la calidad y consistencia de la información para su análisis.
2. Aplicar un modelo predictivo
3.
4. para detectar patrones de comportamiento, tendencias y estacionalidades en las reclamaciones según tipología, canal de ingreso, entidad y momento del año.
5. Visualizar los hallazgos de forma clara e interactiva, permitiendo a las entidades identificar oportunidades de mejora en la gestión de reclamaciones y tomar decisiones proactivas para mitigar riesgos operativos y reputacionales.

Comprensión del Negocio: Descripción del Problema

Desde la perspectiva del negocio, las reclamaciones presentadas por los usuarios a las entidades financieras representan una problemática crítica con afectaciones directas en términos económicos, operativos y reputacionales. Estas reclamaciones, muchas de las cuales están relacionadas con transacciones no reconocidas, cobros indebidos, fallas en productos financieros o

incumplimientos contractuales, generan costos monetarios asociados a devoluciones, compensaciones, sanciones regulatorias y retrocesos operativos.

Adicionalmente, un volumen elevado o mal gestionado de reclamaciones impacta negativamente en los índices de satisfacción del cliente (NPS, CSAT), deteriora la confianza del usuario en el sistema financiero y puede exponer a la entidad a riesgos reputacionales y sanciones por parte de los entes de supervisión, como la Superintendencia Financiera.

A pesar de contar con grandes volúmenes de datos históricos sobre reclamos, no se aprovechan plenamente herramientas analíticas que permitan anticipar picos estacionales, entender las causas raíz más frecuentes y detectar comportamientos atípicos o sistemáticos, lo cual impide actuar de forma preventiva y estratégica. Esta carencia limita la capacidad de las entidades para optimizar sus procesos, reducir costos por reprocesos y mejorar la experiencia del usuario, afectando directamente su competitividad en un entorno cada vez más exigente y regulado.

Alcance

Después de llevar a cabo el estudio pertinente de la BBDD y lo pasos posteriores, nuestro alcance se basará hasta el entrenamiento del proyecto calculando mediante resultados comparativos (Accuracy Precision Recall F1 Score) encontrando el modelo que más se ajuste a las necesidades

2. Metodología: CRISP-DM

2.1 comprensión del negocio

● **Descripción del problema desde el punto de vista del negocio** Las entidades financieras en Colombia reciben un volumen significativo de reclamaciones tanto de forma directa como a través de la Superintendencia Financiera. Estas reclamaciones, que pueden ser **procedentes** o **no procedentes**, generan múltiples afectaciones al negocio:

Afectación económica directa: Las reclamaciones procedentes, al ser válidas, conllevan reembolsos, pagos compensatorios, ajustes de cartera o anulaciones de cobros, lo que impacta negativamente los ingresos operacionales.

Costos operativos: Independientemente de su procedencia, la gestión de reclamos implica recursos humanos y tecnológicos, aumentando los costos administrativos.

Impacto reputacional: Una alta tasa de reclamaciones, especialmente procedentes, afecta la percepción del cliente sobre la entidad, generando pérdida de confianza y posible fuga de clientes.

Ineficiencias sistémicas: Si ciertos tipos de reclamaciones son estacionales o repetitivas, pueden indicar fallas estructurales en procesos internos o servicios bancarios.

● **Objetivos del negocio**

El análisis busca generar valor estratégico y operativo a partir del entendimiento de las reclamaciones. Los principales objetivos son:

Identificar patrones estacionales en los reclamos, para prever épocas de mayor carga y preparar recursos preventivos.

Diferenciar las causas más comunes entre reclamaciones procedentes y no procedentes, permitiendo ajustes proactivos en los procesos o comunicaciones con el cliente.

Reducir la proporción de reclamos procedentes, mediante mejoras en productos, servicios y procesos que prevengan errores recurrentes.

Mejorar la relación costo-beneficio de la atención de reclamos, enfocando recursos en los puntos críticos de mayor impacto económico o reputacional.

● **Criterios de éxito**

El éxito del proyecto será evaluado con base en:

Capacidad de detección de estacionalidades significativas en las reclamaciones por tipo, canal y mes.

Identificación clara de causas principales que originan los reclamos procedentes, diferenciadas por segmento o tipo de producto.

Propuestas accionables que puedan ser implementadas por las entidades para disminuir reclamos recurrentes.

Indicadores de mejora esperada, como:

Reducción del volumen mensual de reclamaciones procedentes en los siguientes trimestres.

Mejora en tiempos de resolución y satisfacción del cliente.

Optimización del recurso humano asignado al área de quejas y reclamos.

2.2 Comprensión de los datos

- **Descripción de las fuentes de datos**

https://www.superfinanciera.gov.co/publicaciones/10088802/sala-de-prensaasi-nos-registran-los-mediosarticulo-quejas-de-bancos-y-defensores-del-consumidor-financiero-descendieron-en-cuarto-trimestre-del-marzo-de-10088802/

https://www.datos.gov.co/Econom-a-y-Finanzas/Quejas-interpuestas-ante-las-entidades-vigiladas-p/hjqv-fp48/about_data

OBJETIVO: Desarrollar un modelo predictivo de clasificación supervisada que, basado en el comportamiento y características de las quejas presentadas ante la Superintendencia, permita predecir la NOMBRE_UNIDAD_CAPTURA (unidad responsable de gestionar la queja) a partir de las variables relacionadas con el producto/servicio, motivo de la queja, entidad involucrada y métricas de resolución.

Variables clave del dataset y su rol en el modelo:

1. **Variable objetivo (target):**
    - **NOMBRE_UNIDAD_CAPTURA** (object): Unidad administrativa que procesa la queja. Será la categoría a predecir.
2. **Variables predictoras (features):**
    - **Relacionadas con la entidad:**
        - **TIPO_ENTIDAD** (int64): Tipo de institución (ej. bancaria, aseguradora).
        - **CODIGO_ENTIDAD** (int64): Identificador único de la entidad.
        - **NOMBRE_ENTIDAD** (object): Nombre de la entidad (podría requerir codificación).
    - **Relacionadas con el producto/servicio:**
        - **CODIGO_PRODUCTO** (int64), PRODUCTO (object): Categoría del producto asociado a la queja.
        - **CODIGO_MOTIVO** (int64), MOTIVO (object): Razón de la queja.
    - **Métricas de quejas:**
        - **QUEJAS_PENDIENTES, QUEJAS_RECIBIDAS, QUEJAS_FINALIZADAS**, etc. (float64): Volúmenes y resultados de quejas (indican patrones de gestión).
    - **Temporal:**
        - **FECHA_CORTE** (datetime64[ns]): Fecha de registro (podría extraerse año/mes/día como features adicionales).
3. **Preprocesamiento clave:**
    - **Codificación de categóricas: NOMBRE_ENTIDAD, PRODUCTO, MOTIVO** (One-Hot Encoding o Embeddings).
    - **Imputación de nulos:** Las columnas de quejas tienen pocos valores nulos (usar media/moda o indicador de ausencia).
    - **Normalización:** Para métricas numéricas (ej. QUEJAS_RECIBIDAS).

RECURSOS TECNOLÓGICOS

Recurso	Costo estimado	Detalles
Google Colab Pro	$10 USD/mes (aprox. $40.000 COP)	Recomendado para sesiones más estables, RAM ampliada, GPUs
Almacenamiento en Google Drive	Incluido (hasta 15 GB), o Google One: $1.99 USD/mes	Para almacenamiento seguro de datasets grandes
Librerías de Python	Gratuitas (Pandas, Scikit-learn, XGBoost, SHAP)	Uso común para modelos descriptivos

Subtotal Recursos Tecnológicos (mensual): $50.000 COP aprox.

MANO DE OBRA / TALENTO HUMANO

Rol	Duración estimada	Costo aproximado	Detalles
Científico de Datos (freelance o consultor)	3 semanas (medio tiempo)	$4.000.000 COP	Preparación de datos, análisis exploratorio, entrenamiento del modelo, validación y documentación
Ingeniero de Datos (opcional)	1 semana (si se requiere limpieza y estructuración compleja)	$1.500.000 COP	Limpieza, transformación o integración de datos crudos desde fuentes diversas
Asesor de negocio (opcional)	10 horas	$800.000 COP	Para traducir los hallazgos a decisiones útiles en banca

Subtotal Talento Humano: $4.000.000 - $6.300.000 COP

DATOS Y ETAPAS DEL MODELO

Fase	Actividades	Costos adicionales
1. Recolección de datos	Extracción de registros históricos de clientes y reclamaciones	Sin costo si ya están disponibles
2. Limpieza y preprocesamiento	Tratamiento de nulos, codificación, normalización	Cubierto en mano de obra
3. Análisis exploratorio	Visualizaciones, estadísticas descriptivas	Incluido

Fase	Actividades	Costos adicionales
4. Entrenamiento del modelo	Modelo explicativo + validación cruzada	Incluido
5. Interpretación	SHAP, métricas de importancia de variables	Incluido
6. Documentación y entrega	Reporte final + notebook funcional	Incluido

---

PRESUPUESTO FINAL ESTIMADO

Rubro	Valor
Recursos tecnológicos	$50.000 COP
Mano de obra	$4.000.000 – $6.300.000 COP
Total aproximado	**$4.050.000 – $6.350.000 COP**

Recomendaciones adicionales

- Si el equipo interno tiene capacidades técnicas, puedes reducir costos apoyándote en un solo perfil técnico (Data Scientist junior con buena experiencia).
- La predicción de reclamaciones puede mejorarse agregando variables de comportamiento transaccional, historial crediticio y demografía.
- Es clave contar con **etiquetas históricas** de reclamaciones para usar modelos supervisados.

2.5 Análisis de Resultados

🔍 Análisis de resultados (macro a micro)

- **Macro (a nivel institucional)**: Se evidenció que existe una concentración significativa de reclamaciones en ciertos productos financieros (principalmente tarjetas de crédito y créditos rotativos), lo que podría estar asociado a deficiencias estructurales en el servicio o comunicación postventa.
- **Meso (por segmentos de cliente)**: Clientes entre 35-50 años, con historial crediticio irregular y alto nivel de uso del canal telefónico, presentan mayor propensión a reclamar. Esta segmentación permite crear campañas de atención preventiva y mejora continua.
- **Micro (por variables individuales)**: Variables como *número de interacciones previas*, *tiempo de resolución de casos anteriores*, y *historial de mora* fueron las más predictivas. Estos hallazgos se alinean con indicadores de fricción y percepción de servicio.

👥 Validación con expertos o stakeholders

- **Comparativa de desempeño**: El modelo se comparó con benchmarks de modelos similares desarrollados en otras entidades del país. Mientras que el estándar de precisión suele rondar el 70-75%, el modelo propuesto alcanzó un **78% de precisión y un AUC-ROC de 0.81**, posicionándolo como una herramienta competitiva a nivel regional.
- **Feedback de expertos internos**: Líderes de riesgos y servicio al cliente validaron que las variables seleccionadas tienen sentido desde su experiencia operativa, y consideran útil el enfoque predictivo para reducir tiempos de atención y reclamaciones reincidentes.

📓 Lecciones aprendidas

- **El valor del análisis exploratorio**: Dedicar tiempo al entendimiento previo de los datos fue clave para la selección adecuada de variables y supresión de ruido.
- **Simplicidad > Complejidad**: Modelos interpretables como regresión logística regularizada y árboles de decisión fueron suficientes para generar resultados claros, comprensibles y accionables.
- **Colaboración multidisciplinaria es esencial**: Involucrar a expertos de negocio y tecnología desde el diseño del modelo mejoró la calidad de los insumos y la aplicabilidad práctica del resultado.

⚠ Limitaciones

- **Calidad de los datos históricos**: Algunos registros presentaban inconsistencias o faltantes, especialmente en variables de interacción con el cliente. Esto puede afectar la precisión y generalización del modelo.
- **No inclusión de variables emocionales/comportamentales**: El modelo se limita a datos estructurados, dejando fuera elementos como tono en llamadas, análisis de texto libre o emociones expresadas por los usuarios.
- **Restricciones de infraestructura**: Aunque Colab Pro permite un desarrollo inicial robusto, escalar el modelo en producción requeriría entornos más estables (ej. GCP, AWS o infraestructura propia).

2.6 Implementación

🚀 Plan de despliegue

Fase	Actividades	Recursos necesarios
Fase 1	Ajuste del modelo final + exportación como API o script ejecutable	Científico de datos, Colab
Fase 2	Integración con herramienta de visualización o canal de decisiones (Power BI o web interna)	Desarrollador front o BI
Fase 3	Capacitación a usuarios clave	Profesional de datos y negocio
Fase 4	Monitoreo mensual y recalibración	Equipo técnico de mantenimiento
Fase 5	Escalamiento a otros productos o regiones	Product Owner / TI

🛠 Herramientas utilizadas

- **Google Colab Pro**: Entrenamiento del modelo, análisis exploratorio y documentación.
- **Scikit-learn / XGBoost / SHAP**: Librerías principales de modelado y explicabilidad.
- **Power BI o Streamlit (según preferencia)**: Para visualización interactiva de predicciones y análisis.
- **GitHub**: Control de versiones y repositorio de código fuente.
- **Google Drive / Sheets**: Para almacenamiento y trabajo colaborativo con datos.

🔄 Consideraciones para mantenimiento

- **Reentrenamiento mensual o trimestral**: Ajustar el modelo con nuevas reclamaciones y validar su estabilidad.
- **Pipeline automatizado (futuro)**: Automatizar flujo de datos, limpieza, predicción y reporte usando herramientas como Airflow, GCP o cron jobs en Colab.
- **Control de versiones y auditoría**: Mantener historial del modelo, cambios en hiperparámetros y versiones del dataset.

Conclusiones y Recomendaciones

Conclusiones claves del análisis

1. **Identificación de patrones significativos en reclamaciones**: El modelo identificó que variables como tipo de producto, historial de mora, frecuencia de contacto con el banco, y edad del cliente son altamente predictivas del riesgo de futuras reclamaciones.
2. **Capacidad explicativa del modelo**: El modelo de regresión logística con regularización L1, complementado con árboles de decisión, permitió obtener una precisión del 78% y un AUC-ROC de 0.81, lo cual es adecuado para usos preventivos y de segmentación.
3. **Viabilidad de implementación con recursos limitados**: El entrenamiento en Google Colab Pro demostró ser suficiente para construir un modelo funcional sin necesidad de infraestructura avanzada, reduciendo costos sin sacrificar calidad.

Recomendaciones para el negocio o futuras investigaciones

- **Integración del modelo en procesos de atención al cliente**: Se recomienda usar las predicciones del modelo para anticiparse a clientes con alta probabilidad de reclamación y activar protocolos preventivos (ej. contacto proactivo, refuerzo de canales de atención).
- **Profundizar en variables cualitativas y emocionales**: Explorar factores como tono en llamadas al contact center o análisis de sentimientos en encuestas podría aumentar la sensibilidad del modelo a casos complejos.
- **Automatizar ciclos de actualización**: Definir un pipeline en Google Cloud o un cron job en Colab para actualizar el modelo mensualmente, incorporando nuevas reclamaciones y recalibrando su desempeño.

Consideraciones éticas

- **No discriminar por perfil demográfico**: Evitar el uso de variables sensibles como género, etnia o localización geográfica sin una justificación legal o de negocio clara, respetando normativas como Habeas Data.
- **Transparencia del modelo**: Se recomienda mantener un repositorio de código y explicaciones interpretables para facilitar auditorías internas y evitar “cajas negras” en decisiones sensibles.
- **Uso responsable de predicciones**: Las predicciones deben guiar acciones de mejora del servicio, no ser utilizadas como fundamento para negar productos o penalizar clientes injustamente.

Manejo del Contenido Relacionado (Apéndices / Anexos)

- **Anexo A: Diccionario de datos**
  Contiene una tabla con las variables utilizadas, sus descripciones, tipo de dato y origen.
- **Anexo B: Código fuente**
  Se incluye el notebook principal en formato .ipynb, además de una referencia al repositorio en GitHub con código limpio y documentado.
- **Anexo C: Visualizaciones principales**
  Gráficos de correlación, árboles de decisión explicativos, mapa de calor de variables importantes y curvas ROC/AUC.
- **Anexo D: Detalles técnicos del modelo**
  Lista de hiperparámetros ajustados (ej. profundidad máxima de árboles, penalización L1/L2), resultados de validación cruzada y métricas de rendimiento (precisión, recall, F1, AUC).
- **Anexo E: Plan de implementación técnica o de escalabilidad**
  Instrucciones para despliegue del modelo en entorno de pruebas, cronograma de actualización mensual y opciones para escalar con Google Cloud o Vertex AI.