Estudio de Caso sobre Sesgo en Modelos de IA en el Sector Financiero

**Título del estudio del Caso: Sesgo Algorítmico en el Otorgamiento de Créditos: El Caso de Apple Card**

Alumna: Johana Patricia Pineda Zamora

Curso: Al Machine Learning - 3

Contenido

[1. Introducción 2](#_Toc204020539)

[2. Descripción del Caso 2](#_Toc204020540)

[3. Identificación del Sesgo 2](#_Toc204020541)

[4. Análisis de las Causas 2](#_Toc204020542)

[5. Propuestas de Mitigación 3](#_Toc204020543)

[6. Conclusión 4](#_Toc204020544)

[7. Referencias 4](#_Toc204020545)

# 1. Introducción

En la última década, los modelos de inteligencia artificial se han convertido en una herramienta esencial para la toma de decisiones en el sector financiero. Sin embargo, la creciente automatización ha puesto en evidencia diversos casos de sesgo algorítmico. Uno de los casos más notables ocurrió con Apple Card, lanzada por Apple en colaboración con Goldman Sachs, donde se reportaron discrepancias injustificadas en los límites de crédito otorgados a hombres y mujeres.  
  
Este informe presenta un análisis del sesgo identificado, sus posibles causas técnicas y sociales, y una serie de recomendaciones para mitigar estos riesgos en futuras implementaciones de IA en el ámbito financiero.

# 2. Descripción del Caso

Apple Card fue lanzada en 2019 como una tarjeta de crédito respaldada por un algoritmo para decidir el límite de crédito. Varios usuarios, incluido el cofundador de Apple, Steve Wozniak, reportaron que sus esposas recibieron límites de crédito significativamente más bajos, pese a tener ingresos e historiales crediticios similares o mejores.  
  
Ejemplo:  
- Steve Wozniak: Límite 10x mayor que su esposa.  
- Ambos declaraban bienes conjuntos y buen historial financiero.

# 3. Identificación del Sesgo

Se identificó un sesgo de género en la asignación de crédito. La IA parecía sistemáticamente asignar menores líneas de crédito a mujeres.  
  
Tipos de sesgo presentes:  
- Sesgo de entrenamiento: Datos históricos con desigualdades de género.  
- Sesgo de representación: Subrepresentación de mujeres en ciertos segmentos financieros.  
- Sesgo implícito de diseño: Variables correlacionadas con género sin un control adecuado.

# 4. Análisis de las Causas

- Datos históricos contaminados: La IA fue entrenada con datos financieros que ya reflejaban desigualdades estructurales.  
- Falta de transparencia del modelo: Al ser un sistema propietario, no se conocía cómo se tomaban las decisiones.  
- Ausencia de fairness metrics: No se aplicaron métricas de equidad para validar la salida del modelo.  
- Desbalance en el dataset: Menor representación de mujeres con historiales financieros extensos.

# 5. Propuestas de Mitigación

Para reducir o eliminar el sesgo en modelos de inteligencia artificial en el sector financiero, se proponen las siguientes estrategias

**Auditoría de Datos Previos**

Antes de entrenar un modelo, es fundamental realizar una auditoría exhaustiva del conjunto de datos para identificar sesgos históricos y estructurales.

* **Objetivo:** Detectar variables proxy de género, raza u otros atributos sensibles que puedan influir indirectamente en las decisiones.
* **Acciones específicas:**
* Analizar la representatividad de los grupos demográficos.
* Identificar correlaciones peligrosas entre variables sensibles y objetivos del modelo.
* Eliminar o ajustar variables con sesgo explícito o implícito.

**Aplicación de Métricas de Equidad (Fairness Metrics)**

Es necesario evaluar el desempeño del modelo con métricas que midan la equidad entre diferentes grupos poblacionales.

* **Métricas recomendadas:**
* Demographic Parity: Igual probabilidad de aprobación entre grupos.
* Equal Opportunity: Igual tasa de verdaderos positivos para cada grupo.
* Disparate Impact Ratio: Comparación proporcional entre resultados favorables por grupo (debe ser ≥ 0.8).
* **Aplicación:** Estas métricas deben calcularse tanto en el conjunto de validación como en producción.

**Modelos Interpretable usando AI**

Los modelos explicables ayudan a comprender cómo y por qué se toman decisiones individuales o colectivas.

Ventajas:

* Detectan decisiones que dependen de variables problemáticas.
* Aumentan la confianza del usuario y permiten auditorías posteriores.

Técnicas comunes:

* LIME (Local Interpretable Model-agnostic Explanations).
* SHAP (SHapley Additive exPlanations).
* Árboles de decisión o regresión lineal como modelos base por su transparencia.

**Validación Cruzada por Segmento Demográfico**

Además de evaluar el rendimiento global del modelo, se deben analizar los resultados por subgrupos como género, edad o nivel socioeconómico.

* Objetivo: Identificar si el modelo generaliza correctamente en cada segmento.
* Estrategias:
* Crear subconjuntos de prueba por grupo demográfico.
* Comparar precisión, recall, F1-score y fairness metrics entre ellos.

# 

# 6. Conclusión

Este caso demuestra cómo un modelo de IA o Machine Learnign mal entrenado o validado puede amplificar desigualdades preexistentes. La equidad en los sistemas algorítmicos no solo es una cuestión técnica, sino también ética. Mitigar el sesgo no solo mejora la precisión del modelo, sino que también refuerza la confianza del público en los sistemas financieros impulsados por IA.

# 7. Referencias

- Wired. (2019). Apple Card investigated after gender bias allegations  
- Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning  
- IBM. (2020). AI Fairness 360 Toolkit