Clasificación de riesgo crediticio en clientes bancarios

**Autor:** Kevin Gómez Villanueva

**Fecha:** 10 de diciembre de 2024

**Resumen ejecutivo:**

Este informe presenta la implementación de un pipeline de Machine Learning usando los servicios de AWS Sagemaker y AWS Bedrock para la clasificación de riesgo crediticio de clientes bancarios. Se aborda el etiquetado de datos usando Modelos Grandes de Lenguaje (LLM), el entrenamiento de un modelo de Machine Learning usando RandomForest y el despliegue del modelo en un endpoint de Sagemaker para las inferencias.

**Índice**

[1. Introducción 1](#_Toc184727167)

[2. Objetivos 1](#_Toc184727168)

[3. Metodología 2](#_Toc184727169)

## Introducción

La clasificación de riesgo crediticio en clientes bancarios es un proceso analítico que evalúa la probabilidad de que un cliente incumpla con el pago de sus obligaciones financieras. Su objetivo es asignar una calificación que refleje el nivel de riesgo asociado con otorgar crédito a un cliente.

La importancia de este proceso radica en su capacidad para ayudar a las instituciones financieras a gestionar el riesgo de manera eficiente, minimizando pérdidas por incumplimientos y optimizando la asignación de recursos.

# Objetivos

Objetivo general: Determinar el nivel de riesgo crediticio de un cliente a partir de sus datos.

Objetivos específicos:

* Utilizar los servicios de AWS Bedrock para elaborar una descripción del perfil del cliente y el etiquetado automático de datos.
* Entrenar y desplegar un modelo de Machine Learning eficiente usando AWS Sagemaker.

# Metodología

**Base de datos:** Se obtiene la base de datos inicial del archivo CSV credit\_risk\_reto.csv. Esta base de datos contiene 1000 entradas con 9 columnas categóricas y numéricas donde cada entrada representa a un cliente. El detalle de las columnas se presenta a continuación.

* Age: Edad de la persona
* Sex: Sexo de la persona
* Job: ( 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
* Housing: Tipo de alojamiento
* Saving accounts: Tipo de cuenta de ahorro
* Checking account: Tipo de cuenta corriente
* Credit amount: Monto de crédito
* Duration (meses): Tiempo de préstamo
* Purpose: Motivo del préstamo

**Exploración de datos:** Se realiza una revisión del dataset para encontrar datos faltantes, datos incorrectos, cantidad de datos, datos numéricos, datos categóricos.

**Etiquetado de datos:** Se realiza la limpieza de datos en base a lo analizado en la exploración y se procede a utilizar AWS Bedrock para el etiquetado automático de datos.

**Preprocesamiento de datos:** Se acondicionan los datos para ser utilizados en el entrenamiento de modelos de Machine Learning.

**Entrenamiento de modelo:** Se utilizan los datos preprocesados para entrenar diferentes modelos y determinar el más adecuado que será utilizado en producción.

**Despliegue del modelo:** El modelo seleccionado se aloja en un endpoint de AWS Sagemaker para realizar inferencias.

Debido a que el proceso de llevar desde la base de datos hasta un modelo en producción requiere de una construcción por etapas, se plantean 3 etapas.

En la Etapa 1 se desarrolla el análisis de los datos, el etiquetado, preprocesamiento, entrenamiento de modelos candidatos con hiperparametrización y la evaluación.

En la Etapa 2 se lleva el modelo a producción tomando en cuenta que se pueden sumar nuevos datos para mantener el modelo eficiente.

En la Etapa 3 se integra la inferencia con el modelo actualizado y todo el pipeline se mantiene en constante monitoreo.

El diagrama se muestra en la Figura 1.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

**Figura 1.** Diagrama de procesos del pipeline de Machine Learning.

# Desarrollo