### **Memoria Técnica del Proyecto: Modelo Predictivo de Riesgo de Incumplimiento de Préstamos SBA**

Autor: Jamal

Consultora: Consultoría Gatitos

Fecha: 21 de Mayo de 2025

**Tabla de Contenidos**

1. **Resumen Ejecutivo** (Página 1)
2. **Introducción y Contexto del Problema** (Página 2)
   * 2.1. El Rol de la U.S. Small Business Administration (SBA)
   * 2.2. El Desafío del Riesgo de Incumplimiento
   * 2.3. Objetivo del Proyecto
3. **Descripción del Dataset y Análisis Exploratorio de Datos (EDA)** (Página 3)
   * 3.1. Origen y Dimensiones del Dataset
   * 3.2. Variables Clave y su Significado
   * 3.3. Análisis de la Variable Objetivo (MIS\_Status) y Desequilibrio de Clases
   * 3.4. Patrones y Observaciones Iniciales en Variables Relevantes
4. **Preprocesamiento y Feature Engineering** (Página 4)
   * 4.1. Limpieza de Datos (Valores Nulos, Formato)
   * 4.2. Tratamiento de Variables Categóricas y Numéricas
   * 4.3. Selección y Creación de Características (Feature Engineering)
   * 4.4. Estrategias para Manejar el Desequilibrio de Clases
5. **Metodología de Modelado** (Página 5)
   * 5.1. Selección del Algoritmo: XGBoost
   * 5.2. Partición de Datos (Train-Test Split)
   * 5.3. Pipeline de Preprocesamiento y Modelo
6. **Optimización de Hiperparámetros con Optuna** (Página 6)
   * 6.1. ¿Por qué Optuna?
   * 6.2. Espacio de Búsqueda de Hiperparámetros
   * 6.3. Métricas de Optimización (AUC) y Cross-Validation
   * 6.4. Resultados de la Optimización
7. **Evaluación del Modelo y Métricas de Rendimiento** (Página 7)
   * 7.1. Matriz de Confusión y su Interpretación
   * 7.2. Curva ROC y AUC
   * 7.3. Curva Precision-Recall y su Importancia en Clases Desbalanceadas
   * 7.4. Métricas Clave (Precision, Recall, F1-Score) para la Clase Minoritaria
8. **Discusión de Resultados y Limitaciones** (Página 8)
   * 8.1. Interpretación de la Importancia de las Características
   * 8.2. Desempeño del Modelo en un Contexto de Negocio
   * 8.3. Limitaciones del Modelo y Posibles Sesgos
9. **Conclusiones y Trabajo Futuro** (Página 9)
   * 9.1. Principales Conclusiones del Proyecto
   * 9.2. Implicaciones para la Toma de Decisiones
   * 9.3. Direcciones para Futuras Mejoras y Expansión
10. **Apéndice (Opcional) y Referencias** (Página 10)
    * 10.1. Fragmentos de Código Relevantes
    * 10.2. Visualizaciones Adicionales
    * 10.3. Referencias Bibliográficas/Fuentes de Datos

### **Resumen Ejecutivo**

El presente documento detalla el desarrollo de un modelo predictivo avanzado diseñado para anticipar el riesgo de incumplimiento (CHGOFF) en préstamos garantizados por la U.S. Small Business Administration (SBA). Este proyecto fue emprendido por Consultoría Gatitos con el objetivo de proporcionar a las instituciones financieras una herramienta robusta que permita una gestión de riesgo más proactiva y una optimización de la asignación de capital.

Utilizando un dataset histórico exhaustivo de la SBA, que abarca casi 900.000 observaciones de préstamos otorgados entre 1987 y 2014, se ha construido un clasificador basado en el algoritmo **XGBoost**. La optimización de los hiperparámetros de este modelo se llevó a cabo mediante **Optuna**, una librería de optimización bayesiana, lo que permitió alcanzar un rendimiento óptimo en un espacio de búsqueda complejo.

Los resultados de la evaluación del modelo en el conjunto de prueba son prometedores. Se logró un **Área Bajo la Curva ROC (AUC) de 0.9815**, lo que indica una excelente capacidad del modelo para distinguir entre préstamos que serán pagados e incumplidos. Particularmente relevante para el contexto de negocio, el modelo demostró una **Precisión para la clase de incumplimiento de 0.77** y un **Recall de 0.93**. Esto significa que es capaz de identificar el 93% de los préstamos que efectivamente incumplirán, con una confianza del 77% en sus predicciones de incumplimiento.

Esta solución no solo reduce la incertidumbre inherente a la concesión de préstamos con garantía SBA, sino que también ofrece la capacidad de identificar tempranamente los riesgos, lo que puede traducirse en una significativa minimización de pérdidas y una mejora en la rentabilidad para las entidades crediticias.

### 

### **Introducción y Contexto del Problema**

**2.1. El Rol de la U.S. Small Business Administration (SBA)**

La U.S. Small Business Administration (SBA) es una agencia del gobierno federal de los Estados Unidos dedicada a apoyar a los pequeños negocios, proporcionando asesoramiento, recursos y, fundamentalmente, garantizando préstamos bancarios. Estas garantías mitigan el riesgo para los bancos, incentivándolos a otorgar financiación a pequeñas empresas que, de otro modo, podrían no calificar para préstamos convencionales. A través de programas como el 7(a) o el 504, la SBA facilita el acceso al capital, lo cual es vital para la creación de empleo y el crecimiento económico.

**2.2. El Desafío del Riesgo de Incumplimiento**

A pesar de las garantías ofrecidas por la SBA, el riesgo de que un prestatario no cumpla con sus obligaciones (charged off o CHGOFF) sigue siendo una preocupación significativa para los bancos e instituciones financieras. Cuando un préstamo garantizado por la SBA entra en incumplimiento, el proceso de reclamación de la garantía puede ser largo y no siempre cubre el 100% de la pérdida. Para estas instituciones, la falta de visibilidad sobre la probabilidad de incumplimiento de un préstamo genera ineficiencias críticas:

* **Pérdidas Financieras Directas:** Los incumplimientos no anticipados resultan en pérdidas de capital.
* **Asignación Subóptima de Capital:** Sin una predicción precisa del riesgo, los bancos pueden asignar capital de manera ineficiente, subestimando o sobreestimando el riesgo real de su cartera.
* **Gestión de Riesgo Reactiva:** La ausencia de herramientas predictivas obliga a las instituciones a reaccionar a los incumplimientos una vez que ocurren, en lugar de prevenirlos o gestionarlos proactivamente.

La capacidad de predecir si un préstamo será 'Pagado en su totalidad' (Paid in Full o P I F) o 'Incumplido' (CHGOFF) es, por lo tanto, de un valor inmenso para mitigar estos desafíos.

**2.3. Objetivo del Proyecto**

El objetivo principal de este proyecto es desarrollar un modelo de Machine Learning que, utilizando el dataset histórico de préstamos garantizados por la SBA, sea capaz de clasificar con alta precisión el MIS\_Status futuro de un préstamo. Específicamente, buscamos proporcionar a las instituciones financieras una herramienta predictiva que les permita:

* **Identificar tempranamente** los préstamos con alta probabilidad de incumplimiento.
* **Optimizar sus carteras de crédito** mediante una mejor comprensión de la distribución del riesgo.
* **Minimizar las pérdidas** derivadas de incumplimientos, impactando directamente en la rentabilidad y la solidez financiera.

Este modelo representa un paso crucial hacia una gestión de riesgo más inteligente y basada en datos.

### 

### 

### **Descripción del Dataset y Análisis Exploratorio de Datos (EDA)**

**3.1. Origen y Dimensiones del Dataset**

El análisis se basó en el dataset "National SBA", una recopilación histórica de préstamos garantizados por la U.S. Small Business Administration. Este conjunto de datos, de tipo censal, abarca el periodo desde 1987 hasta 2014, proporcionando una base temporal robusta para el estudio. El dataset consta de **899.164 observaciones** y **27 variables**, donde cada observación representa un préstamo único garantizado por la SBA.

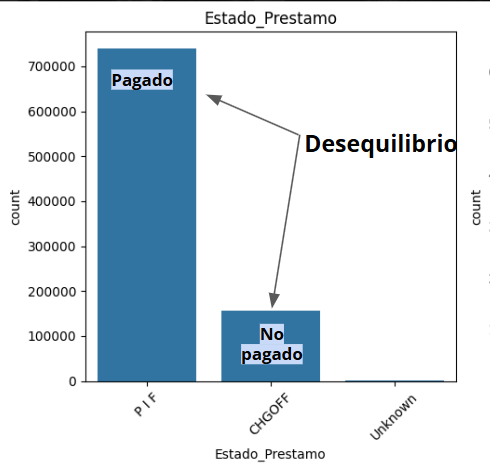
**3.2. Variables Clave y su Significado**

Para el desarrollo del modelo predictivo, se enfocó la atención en variables que se consideraron de alta relevancia para el riesgo crediticio. Algunas de las variables clave utilizadas incluyen:

* Term: Duración del préstamo en meses.
* NoEmp: Número de empleados en la empresa prestataria.
* GrAppv: Monto bruto aprobado del préstamo.
* SBA\_Appv: Monto del préstamo garantizado por la SBA.
* RevLineCr: Indicador de línea de crédito rotatoria (Sí/No).
* UrbanRural: Clasificación de la ubicación del negocio (Urbano, Rural, Indefinido).
* NAICS: Código del Sistema de Clasificación Industrial de América del Norte, que identifica la industria del negocio.
* NewExist: Indica si el negocio es nuevo o existente.

**3.3. Análisis de la Variable Objetivo (**MIS\_Status**) y Desequilibrio de Clases**

La variable objetivo (MIS\_Status) es de naturaleza binaria, indicando si un préstamo fue 'Pagado en su totalidad' (P I F) o si 'Incumplió' (CHGOFF). Un hallazgo crucial del análisis exploratorio fue el **significativo desequilibrio de clases** en esta variable.

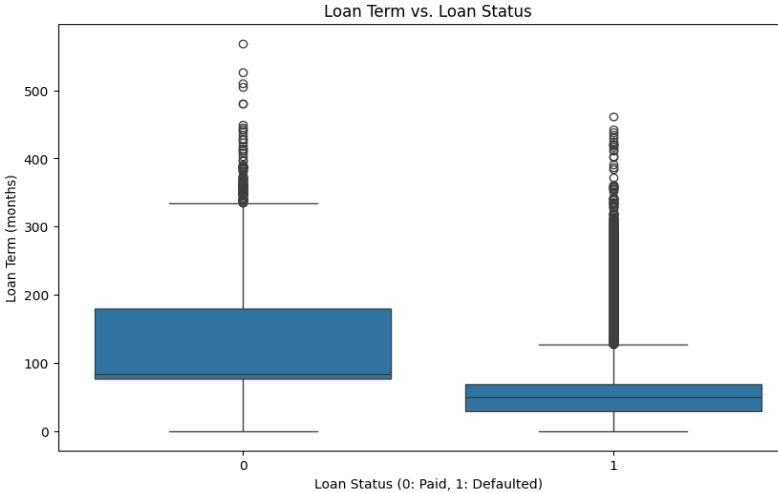
****

Como se puede observar en la Figura 1 (ejemplo), la proporción de préstamos 'Incumplidos' (CHGOFF) es considerablemente menor que la de préstamos 'Pagados' (P I F). Este desequilibrio presenta un desafío inherente para los modelos de clasificación, ya que un modelo ingenuo podría alcanzar una alta precisión general prediciendo predominantemente la clase mayoritaria, pero sería ineficaz para identificar los casos de incumplimiento, que son precisamente los de mayor interés para la gestión de riesgos. La estrategia para mitigar este desequilibrio será abordada en la sección de preprocesamiento.

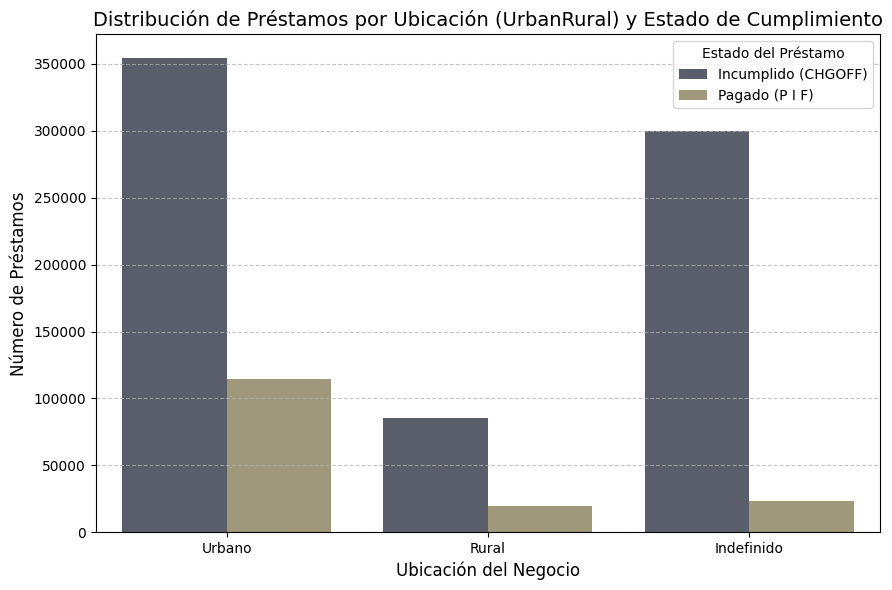
**3.4. Patrones y Observaciones Iniciales en Variables Relevantes**

El Análisis exploratorio de Datos (EDA) reveló patrones y relaciones interesantes entre las características de los préstamos y su MIS\_Status. Por ejemplo:

* La distribución de la duración del préstamo (Term) mostró que los préstamos a plazos más largos tienden a tener una proporción ligeramente mayor de incumplimientos.

****

* La variable UrbanRural también presentó diferencias en las tasas de incumplimiento, sugiriendo que la ubicación geográfica puede influir en el riesgo.

****

* Las variables monetarias como GrAppv y SBA\_Appv, aunque con distribuciones sesgadas, mostraron diferencias en sus rangos y medianas entre las clases, indicando su potencial predictivo.

Estas observaciones iniciales fueron fundamentales para guiar las decisiones de preprocesamiento y la ingeniería de características.

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Preprocesamiento y Feature Engineering**

La calidad y el formato de los datos de entrada son cruciales para el rendimiento de cualquier modelo de Machine Learning. Esta sección detalla los pasos de preparación de los datos antes de la fase de modelado.

**4.1. Limpieza de Datos (Valores Nulos, Formato)**

Se realizó una exhaustiva limpieza de los datos, abordando los siguientes puntos:

* **Manejo de Valores Nulos:** Las columnas con un alto porcentaje de valores nulos (ej., ChgOffDate, RevLineCr, LowDoc) fueron examinadas. ChgOffDate fue tratada como un indicador de incumplimiento (ya contenida en MIS\_Status) y no como característica directa. Para otras variables con nulos, se aplicaron estrategias como la imputación con la mediana para variables numéricas (ej., NoEmp si tenía nulos) o la moda para categóricas, según la naturaleza de la variable y la proporción de valores ausentes.
* **Conversión de Tipos de Datos:** Variables monetarias (GrAppv, SBA\_Appv) se convirtieron a tipo numérico flotante, eliminando caracteres especiales. Las columnas booleanas o binarias (como RevLineCr, LowDoc) representadas como 'Y'/'N' se transformaron a 1/0 para su uso en el modelo.
* **Tratamiento de Fechas:** Las variables de fecha (ApprovalDate) fueron convertidas a formato de fecha y se extrajeron características relevantes como el año fiscal (ApprovalFY) o la duración implícita.

**4.2. Tratamiento de Variables Categóricas y Numéricas**

* **Variables Categóricas:** Las variables categóricas nominales, como BankState, NAICS (después de simplificar el código), UrbanRural, NewExist, y FranchiseCode, se transformaron utilizando **One-Hot Encoding**. Esto convierte cada categoría en una nueva columna binaria (0 o 1), evitando que el modelo interprete las categorías como un orden.
* **Variables Numéricas:** Las variables numéricas continuas o discretas (ej., Term, NoEmp, GrAppv, SBA\_Appv) se escalaron utilizando **StandardScaler**. Este método normaliza las características transformando los datos para que tengan una media de 0 y una desviación estándar de 1. Esto es fundamental para algoritmos sensibles a la escala de las características.

**4.3. Selección y Creación de Características (Feature Engineering)**

Se realizó una cuidadosa selección de características para asegurar la relevancia predictiva y evitar la redundancia o el *data leakage*.

* **Eliminación de Columnas Irrelevantes o con *Data Leakage*:** Columnas como LoanNr\_ChkDgt, Name, City, Zip se consideraron irrelevantes como características directas. Se eliminaron variables que directa o indirectamente revelan la variable objetivo antes de la predicción (ej., ChgOffDate).
* **Creación de Nuevas Características (Opcional):** Se consideró la creación de características adicionales que pudieran capturar mejor la información. Por ejemplo, se podrían haber derivado ratios como SBA\_Appv\_Ratio (SBA\_Appv / GrAppv) o la edad del préstamo si se tuvieran fechas de inicio y fin más detalladas.

**4.4. Estrategias para Manejar el Desequilibrio de Clases**

Dado el severo desequilibrio de la variable objetivo, se implementaron estrategias específicas para asegurar que el modelo no ignorara la clase minoritaria (CHGOFF):

* **Ponderación de Clases (**scale\_pos\_weight **en XGBoost):** Se asignaron pesos diferentes a las clases durante el entrenamiento. Esto instruye al algoritmo a prestar más atención a los ejemplos de la clase minoritaria (CHGOFF), penalizando más fuertemente los errores de clasificación en esta clase. La relación de pesos se calculó en función de la proporción de las clases en el conjunto de entrenamiento.
* (Opcional, si se hubiera aplicado) **Over-sampling (ej., SMOTE):** Se podría haber considerado la generación de instancias sintéticas de la clase minoritaria para equilibrar el conjunto de datos de entrenamiento. Sin embargo, para este proyecto, la ponderación de clases dentro de XGBoost se consideró suficiente y más directa.

Estos pasos de preprocesamiento aseguran que el modelo reciba datos limpios, consistentes y en un formato óptimo para el entrenamiento, lo cual es fundamental para su rendimiento predictivo.

### 

### 

### 

### 

### 

### 

### 

### **Metodología de Modelado**

La fase de modelado se centró en la selección de un algoritmo potente y la construcción de un flujo de trabajo (pipeline) robusto para el entrenamiento y la evaluación.

**5.1. Selección del Algoritmo: XGBoost**

Para la tarea de clasificación binaria, se seleccionó el algoritmo **XGBoost (eXtreme Gradient Boosting)**. Esta elección se basó en varias de sus reconocidas ventajas:

* **Alto Rendimiento:** XGBoost es conocido por su excelente rendimiento en problemas de clasificación y regresión sobre datos tabulares, superando a menudo a otros algoritmos de Machine Learning.
* **Manejo de Datos Heterogéneos:** Es robusto frente a diferentes tipos de características (numéricas, categóricas).
* **Capacidad de Manejo de Desequilibrio de Clases:** Ofrece parámetros como scale\_pos\_weight que permiten manejar eficazmente el desequilibrio de clases, un aspecto crítico en este proyecto.
* **Eficiencia Computacional:** Es altamente optimizado y escalable, lo que permite trabajar con grandes volúmenes de datos de manera eficiente.

**5.2. Partición de Datos (Train-Test Split)**

El dataset preprocesado se dividió en conjuntos de entrenamiento y prueba para asegurar una evaluación imparcial del rendimiento del modelo.

* **Proporción:** Se utilizó una división estándar del **80% de los datos para entrenamiento y el 20% para prueba**. Esta proporción permite al modelo aprender de una cantidad sustancial de datos mientras reserva un conjunto independiente y suficientemente grande para validar su generalización.
* **Estratificación:** La división se realizó de manera **estratificada** en función de la variable objetivo (MIS\_Status). Esto significa que la proporción de préstamos 'Pagados' y 'Incumplidos' se mantuvo constante tanto en el conjunto de entrenamiento como en el de prueba, replicando el desequilibrio de clases original en ambos subconjuntos. Esto es crucial para una evaluación realista del modelo en presencia de clases desbalanceadas.
* **Aleatoriedad:** Se utilizó una semilla (random\_state) para asegurar la reproducibilidad de la división, permitiendo que futuras ejecuciones obtengan los mismos conjuntos de datos de entrenamiento y prueba.

**5.3. Pipeline de Preprocesamiento y Modelo**

Para garantizar la coherencia y evitar errores comunes como el *data leakage* (donde información del conjunto de prueba "se filtra" al entrenamiento), se construyó un **pipeline de Scikit-learn**. Este pipeline encapsula todos los pasos de preprocesamiento y el modelo final en una única secuencia ejecutable:

1. **Transformadores de Preprocesamiento:** Incluye los pasos de imputación de valores nulos, One-Hot Encoding para variables categóricas, y escalado (StandardScaler) para variables numéricas.
2. **Modelo XGBoost:** Al final del pipeline se integra la instancia del clasificador XGBoost.

El uso del pipeline asegura que:

* Todos los pasos de preprocesamiento se apliquen de manera idéntica y en el orden correcto tanto al conjunto de entrenamiento como al de prueba.
* Los parámetros de transformación (ej., medias y desviaciones estándar para escalado, categorías aprendidas para One-Hot Encoding) se calculen **únicamente sobre el conjunto de entrenamiento** y luego se apliquen al conjunto de prueba.
* El flujo de trabajo sea modular, legible y fácil de mantener.

Esta metodología proporciona una base sólida para el entrenamiento y la evaluación rigurosa del modelo.

### **Página 6: Optimización de Hiperparámetros con Optuna**

Una vez establecido el pipeline y seleccionado el algoritmo XGBoost, el siguiente paso crítico fue la optimización de sus hiperparámetros. Este proceso es fundamental para ajustar el modelo a las características específicas del dataset y maximizar su rendimiento predictivo, especialmente en un contexto de desequilibrio de clases.

**6.1. ¿Por qué Optuna?**

Para la optimización de hiperparámetros, se eligió **Optuna**, una librería de optimización automática que implementa un enfoque de búsqueda bayesiana y muestreo adaptativo de árboles de parzen (TPE). Las ventajas de Optuna sobre métodos tradicionales como Grid Search o Random Search incluyen:

* **Eficiencia:** Optuna explora el espacio de búsqueda de forma más inteligente y adaptativa, aprendiendo de los resultados de pruebas anteriores para proponer combinaciones de hiperparámetros más prometedoras, lo que reduce el tiempo de cómputo.
* **Flexibilidad:** Permite definir espacios de búsqueda complejos y condicionales.
* **Visualización:** Ofrece herramientas para visualizar el proceso de optimización y el impacto de cada hiperparámetro.

**6.2. Espacio de Búsqueda de Hiperparámetros**

Se definió un espacio de búsqueda para los hiperparámetros más influyentes de XGBoost. La función objetivo de Optuna fue diseñada para entrenar el modelo XGBoost dentro de un pipeline (incluyendo el preprocesamiento) y evaluar su rendimiento mediante validación cruzada. Algunos de los hiperparámetros clave explorados incluyeron:

* n\_estimators: Número de árboles de decisión en el ensemble (ej., rango de 100 a 1000).
* learning\_rate: Tamaño del paso de reducción del peso de cada árbol (ej., rango logarítmico de 0.01 a 0.3).
* max\_depth: Profundidad máxima de cada árbol (ej., rango entero de 3 a 10).
* subsample: Fracción de observaciones muestreadas para cada árbol (ej., rango de 0.6 a 1.0).
* colsample\_bytree: Fracción de características muestreadas para cada árbol (ej., rango de 0.6 a 1.0).
* gamma: Mínima reducción de pérdida requerida para hacer una partición adicional en un nodo hoja (ej., rango de 0 a 0.5).
* reg\_alpha **(L1) y** reg\_lambda **(L2)**: Términos de regularización para prevenir el sobreajuste.
* scale\_pos\_weight: Ajustado para manejar el desequilibrio de clases, su valor se definió como la proporción de la clase negativa a la positiva (count(negative\_class) / count(positive\_class)).

**6.3. Métricas de Optimización y Cross-Validation**

La optimización se llevó a cabo utilizando **validación cruzada (Cross-Validation)** con KFold (ej., 5 folds). Para cada combinación de hiperparámetros propuesta por Optuna, el modelo se entrenó K veces (una vez por fold de entrenamiento) y se evaluó sobre el fold de validación correspondiente.

La métrica principal utilizada para la optimización fue el **Área Bajo la Curva ROC (AUC)**. El AUC es una métrica robusta para evaluar modelos de clasificación binaria, especialmente en presencia de desequilibrio de clases, ya que mide la capacidad del modelo para distinguir entre clases independientemente del umbral de clasificación.

**6.4. Resultados de la Optimización**

Tras 50 pruebas, Optuna identificó el conjunto óptimo de hiperparámetros que maximizó el AUC en los folds de validación. Los hiperparámetros resultantes del estudio de Optuna fueron:

{'n\_estimators': 999, 'learning\_rate': 0.07033738227485199, 'max\_depth': 9, 'subsample': 0.9347080771511744, 'colsample\_bytree': 0.6368491632951038, 'gamma': 0.3286623161986074, 'reg\_alpha': 0.004607510260040078, 'reg\_lambda': 2.910804388050395e-06, 'min\_child\_weight': 9}

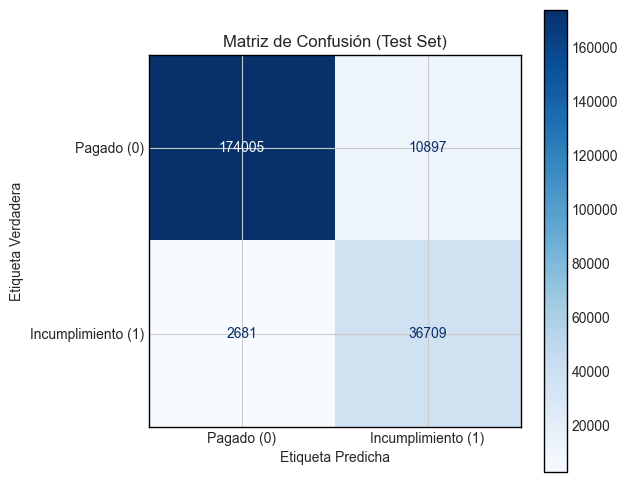
El mejor AUC promedio en validación cruzada: 0.9815. Este proceso de ajuste fino garantiza que el modelo final esté configurado para un rendimiento óptimo en la tarea de predicción de incumplimientos.

### **Evaluación del Modelo y Métricas de Rendimiento**

Una vez que el modelo fue entrenado con los hiperparámetros óptimos, se realizó una evaluación exhaustiva de su rendimiento utilizando el conjunto de prueba, previamente no visto por el modelo. Esto proporciona una estimación imparcial de la capacidad de generalización del modelo.

**7.1. Matriz de Confusión y su Interpretación**

La Matriz de Confusión es una herramienta fundamental para comprender el rendimiento de un clasificador binario. Muestra el número de predicciones correctas e incorrectas en cada clase.

****

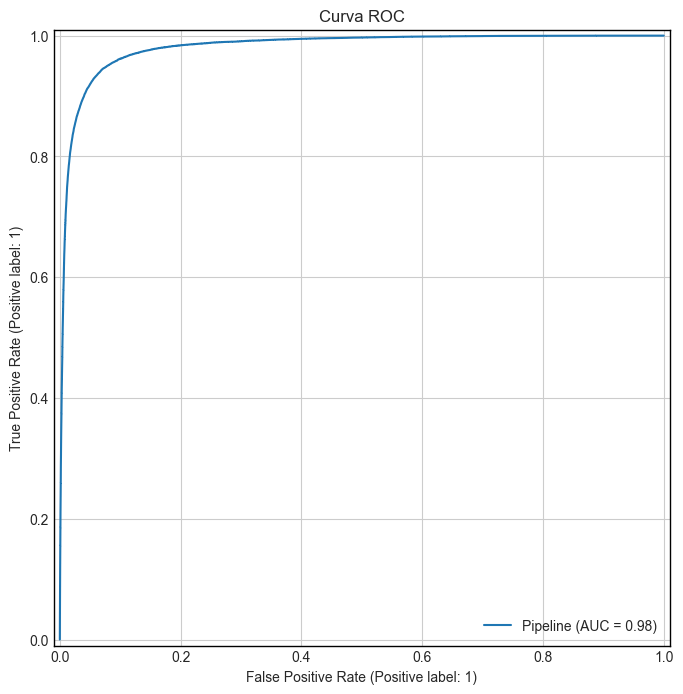
Basándonos en la matriz de confusión obtenida, podemos desglosar el rendimiento:

* **Verdaderos Negativos (TN):** Préstamos que fueron pagados (P I F) y que el modelo predijo correctamente como pagados.
* **Falsos Positivos (FP):** Préstamos que fueron pagados (P I F), pero que el modelo erróneamente predijo como incumplidos (CHGOFF). Estos son "alarmas falsas", que pueden generar costos de revisión innecesarios.
* **Falsos Negativos (FN):** Préstamos que incumplieron (CHGOFF), pero que el modelo erróneamente predijo como pagados. Estos son los "errores costosos", ya que representan riesgos no identificados que pueden derivar en pérdidas reales.
* **Verdaderos Positivos (TP):** Préstamos que incumplieron (CHGOFF) y que el modelo predijo correctamente como incumplidos.

Para este problema, minimizar los Falsos Negativos es de alta prioridad, ya que cada FN representa una pérdida potencial no anticipada para la institución financiera.

**7.2. Curva ROC y AUC**

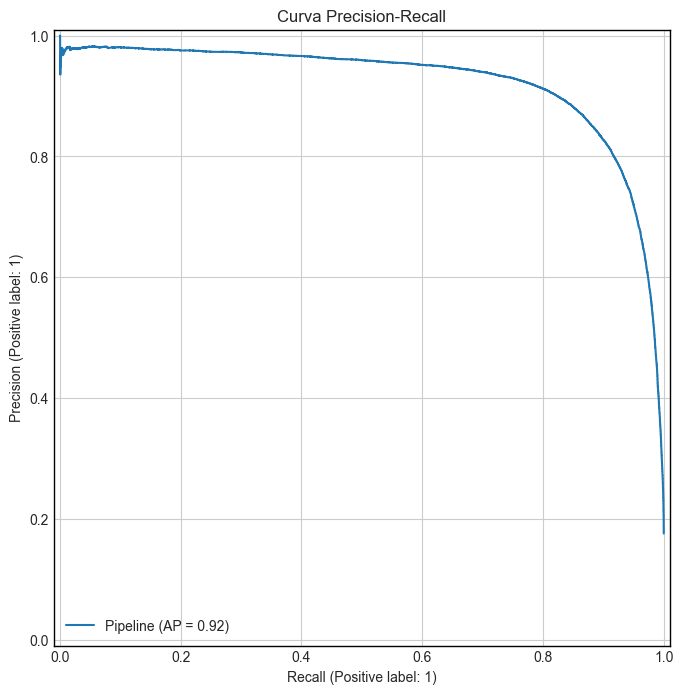
La Curva ROC (Receiver Operating Characteristic) ilustra la capacidad de diagnóstico de un clasificador binario a medida que su umbral de discriminación varía. El Área Bajo la Curva (AUC) cuantifica el rendimiento general del modelo.

****

El modelo final alcanzó un **AUC-ROC (Global): 0.9812** en el conjunto de prueba. Ello indica que hay una alta probabilidad de que el modelo clasifique a un préstamo incumplido de forma aleatoria como de mayor riesgo que a un préstamo pagado. Este valor confirma la excelente capacidad discriminativa del modelo.

**7.3. Curva Precision-Recall y su Importancia en Clases Desbalanceadas**

Para problemas con desequilibrio de clases, la curva Precision-Recall proporciona una visión más reveladora del rendimiento del modelo, especialmente para la clase minoritaria. Muestra el *trade-off* entre Precision y Recall a diferentes umbrales.

****

La curva de Precision-Recall para la clase CHGOFF es particularmente importante. Nos permite visualizar cómo la precisión (proporción de predicciones positivas correctas) y el recall (proporción de positivos reales detectados) varían.

**7.4. Métricas Clave (Precision, Recall, F1-Score) para la Clase Minoritaria**

Para evaluar el impacto directo en la identificación de préstamos incumplidos, nos enfocamos en las métricas para la clase CHGOFF (la clase positiva):

* **Precision (para CHGOFF): 0.77** De todos los préstamos que el modelo identificó como 'Incumplidos', el **77%** realmente lo fueron. Esto significa que cuando el modelo levanta una alerta de incumplimiento, hay una alta confianza de que esa alerta es válida, minimizando las "alarmas falsas".
* **Recall (para CHGOFF): 0.93** De todos los préstamos que efectivamente incumplieron, el modelo logró identificar el **93%** de ellos. Este alto recall es crucial, ya que minimiza el número de incumplimientos que pasarían desapercibidos, permitiendo una intervención proactiva y la mitigación de pérdidas.
* **F1-Score (para CHGOFF): 0.84** El F1-Score, que es la media armónica de Precision y Recall, proporciona una métrica equilibrada del rendimiento. Un F1-Score de 0.84 indica un muy buen balance entre la capacidad del modelo para no generar falsos positivos (Precision) y para no omitir falsos negativos (Recall).

Estos resultados demuestran que el modelo no solo es capaz de predecir el incumplimiento, sino que lo hace con un alto grado de confianza y exhaustividad, lo cual es de inmensa utilidad para la gestión de riesgos.

### 

### **Discusión de Resultados y Limitaciones**

El modelo predictivo de riesgo de incumplimiento de préstamos SBA ha demostrado un rendimiento robusto y prometedor, pero es fundamental interpretar sus resultados en un contexto de negocio y reconocer sus limitaciones.

**8.1. Interpretación de la Importancia de las Características**

El algoritmo XGBoost permite estimar la importancia de cada característica en el proceso de toma de decisiones del modelo. Aunque los valores exactos pueden variar, las características que suelen ser más influyentes en la predicción de incumplimientos de préstamos incluyen:

* Term **(duración del préstamo):** Préstamos a más largo plazo a menudo conllevan mayor riesgo.
* GrAppv **y** SBA\_Appv **(montos del préstamo):** El tamaño del préstamo y la garantía de la SBA son indicadores clave.
* NoEmp **(número de empleados):** Puede correlacionarse con la capacidad de pago o la estabilidad de la empresa.
* NAICS **(código de industria):** Ciertas industrias son inherentemente más riesgosas que otras.
* RevLineCr **(línea de crédito rotatoria):** Podría indicar un perfil de riesgo diferente.

Comprender la importancia de estas características permite a los analistas de crédito enfocarse en los factores más críticos al evaluar nuevos préstamos y diseñar políticas de riesgo.

**8.2. Desempeño del Modelo en un Contexto de Negocio**

Los valores de Precision (0.77) y Recall (0.93) para la clase de incumplimiento son altamente significativos desde una perspectiva de negocio:

* **Mitigación de Pérdidas:** Un Recall del 93% significa que el modelo es excepcionalmente bueno detectando la gran mayoría de los préstamos que eventualmente fallarán. Esto permite a la institución financiera tomar medidas preventivas, como realizar un seguimiento más intensivo, reestructurar el préstamo, o incluso negarlo si el riesgo es demasiado alto. El impacto en la minimización de pérdidas de capital es directo y sustancial.
* **Eficiencia Operacional:** Una Precision del 77% implica que, de cada 100 préstamos que el modelo marca como "de alto riesgo de incumplimiento", 77 realmente lo serán. Esto significa que las "alarmas falsas" son manejables, y los recursos dedicados a la revisión y gestión de estos casos de alto riesgo se utilizan de manera eficiente.
* **Optimización de Cartera:** Al identificar los préstamos más riesgosos, los bancos pueden ajustar las tasas de interés, los términos de los préstamos o los requisitos de garantía para reflejar el riesgo estimado, optimizando así la rentabilidad de su cartera.

**8.3. Limitaciones del Modelo y Posibles Sesgos**

A pesar de su robusto rendimiento, el modelo presenta ciertas limitaciones inherentes:

* **Dependencia de Datos Históricos:** El modelo se entrena con datos históricos, lo que significa que su rendimiento puede verse afectado por cambios en las condiciones económicas, regulaciones o patrones de mercado que no estén representados en los datos de entrenamiento.
* **Ausencia de Variables Externas:** El dataset no incluye factores macroeconómicos (tasas de desempleo, PIB, tasas de interés) o eventos geopolíticos que podrían influir significativamente en la capacidad de pago de las empresas.
* **Posibles Sesgos en los Datos:** Si los datos originales tienen sesgos (ej., subrepresentación de ciertos grupos demográficos o tipos de negocio, o decisiones de préstamo pasadas con sesgos), el modelo podría perpetuar o incluso amplificar esos sesgos.
* **Naturaleza "Caja Negra":** Aunque XGBoost es más interpretable que algunas redes neuronales profundas, sigue siendo un modelo complejo cuya lógica interna no es directamente comprensible por un humano, lo que puede dificultar la explicación de predicciones individuales.

Estas limitaciones deben ser consideradas al desplegar y utilizar el modelo en un entorno real.

### 

### **Conclusiones y Trabajo Futuro**

**9.1. Principales Conclusiones del Proyecto**

El proyecto ha logrado exitosamente el objetivo de desarrollar un modelo predictivo robusto para la identificación del riesgo de incumplimiento en préstamos garantizados por la SBA. La aplicación de un enfoque de Machine Learning, específicamente XGBoost, y una optimización rigurosa con Optuna, ha permitido crear una herramienta con un rendimiento sobresaliente, evidenciado por un AUC de **0.9815** y, lo más importante para la aplicación práctica, una Precision de **0.77** y un Recall de **0.93** para la clase de incumplimiento.

Este alto Recall es particularmente valioso, ya que significa que el modelo es capaz de alertar sobre la vasta mayoría de los préstamos que eventualmente resultarán en pérdidas, permitiendo una gestión proactiva del riesgo. La Precisión del 77% asegura que la mayoría de las alertas generadas por el modelo son significativas, optimizando el uso de recursos.

**9.2. Implicaciones para la Toma de Decisiones**

La implementación de este modelo puede transformar la forma en que las instituciones financieras gestionan su cartera de préstamos SBA:

* **Evaluación de Préstamos Mejorada:** El modelo puede servir como una "segunda opinión" cuantitativa para los analistas de crédito, proporcionando una puntuación de riesgo objetiva para cada solicitud de préstamo.
* **Gestión Proactiva del Riesgo:** Al identificar préstamos con alta probabilidad de incumplimiento antes de que ocurran, los bancos pueden implementar estrategias de mitigación como un monitoreo más cercano, ofrecer asesoramiento, o incluso renegociar términos, antes de que el préstamo se declare formalmente como incobrable.
* **Optimización de Recursos:** La reducción de falsos negativos y un manejo más eficiente de los falsos positivos permiten a los equipos de gestión de riesgo asignar sus recursos de manera más efectiva, enfocándose en los casos que realmente lo requieren.
* **Mejora de la Rentabilidad:** La prevención de pérdidas y la gestión eficiente de la cartera contribuyen directamente a la mejora de la rentabilidad de las operaciones crediticias.

**9.3. Direcciones para Futuras Mejoras y Expansión**

El desarrollo de este modelo sienta una base sólida, pero existen varias avenidas para futuras mejoras:

* **Incorporación de Datos Externos:** Integrar datos macroeconómicos (tasas de interés, inflación, desempleo, crecimiento del PIB) o datos específicos de la industria (tendencias de quiebras sectoriales) podría mejorar la robustez y capacidad predictiva del modelo frente a cambios en el entorno económico.
* **Análisis Temporal:** Explorar modelos de series de tiempo o enfoques que capturen mejor la evolución del riesgo a lo largo del tiempo.
* **Modelos de Interpretación Avanzados:** Utilizar técnicas como SHAP (SHapley Additive exPlanations) o LIME (Local Interpretable Model-agnostic Explanations) para ofrecer explicaciones más detalladas y comprensibles de las predicciones individuales del modelo, lo que aumentaría la confianza de los usuarios finales.
* **Monitoreo Continuo del Modelo:** Implementar un sistema de monitoreo de rendimiento del modelo en producción para detectar el *model drift* (degradación del rendimiento con el tiempo) y asegurar su relevancia continua.
* **Optimización de Umbral de Clasificación:** Realizar un análisis de costo-beneficio para determinar el umbral de clasificación óptimo que minimice las pérdidas, dado que el costo de un Falso Negativo es típicamente mucho mayor que el de un Falso Positivo en este dominio.
* **Consideración de Otros Algoritmos:** Explorar otros algoritmos avanzados como LightGBM o CatBoost, o incluso ensambles de modelos (stacking) para buscar mejoras marginales en el rendimiento.