Comparación de algoritmos de DL frente a algoritmos de ML clásicos

Caso de Estudio: Predicción del incumplimiento de pago de préstamos bancarios garantizados por la U.S. Small Business Administration

|  |  |  |
| --- | --- | --- |
| Titulación:  Máster Universitario en Inteligencia Artificial  Curso académico  2021 - 2023 | Alumno: Juan Alex Castro Gumiel  D.N.I: 5074733  Director de TFM: José Antonio Lagares Rodríguez | Convocatoria:  Cuarto Periodo |

**Índice**

[1. Introducción 2](#_Toc146416833)

[2. Ámbito de la Aplicación 2](#_Toc146416834)

[3. Estado del Arte 3](#_Toc146416835)

[4. Marco Teórico 9](#_Toc146416836)

[4.1 Conceptos Clave 9](#_Toc146416837)

[4.2 Investigaciones Previas 10](#_Toc146416838)

[4.3 Gap de Investigación 10](#_Toc146416839)

[4.4 Hipótesis de Investigación 10](#_Toc146416840)

[5. Metodología 11](#_Toc146416841)

[6. Objetivos 12](#_Toc146416842)

[6.1 Objetivos Generales 12](#_Toc146416843)

[6.2 Objetivos Específicos 13](#_Toc146416844)

[7. Desarrollo de Modelos de IA 13](#_Toc146416845)

[7.1 Análisis Exploratorio de Datos 13](#_Toc146416846)

[7.2 Preprocesamiento de Datos 13](#_Toc146416847)

[7.3 Modelos de ML: Clasificación 13](#_Toc146416848)

[7.4 Modelos de ML: Agrupación 13](#_Toc146416849)

[7.5 Modelos de Deep Learning 13](#_Toc146416850)

[7.6 Aplicación de MLOps 13](#_Toc146416851)

[7.7 Comparación de Resultados 13](#_Toc146416852)

[8. Resultados 14](#_Toc146416853)

[9. Conclusiones 14](#_Toc146416854)

**Índice de tablas**

[**Tabla 1** Optimal parameters for our algorithm after Gridsearch. 4](#_Toc146488123)

[**Tabla 2** Performance metric for XGBoost algorithm on the loan dataset. 5](#_Toc146488124)

[**Tabla 3** Evaluation metrics comparison of the four techniques. 5](#_Toc146488125)

[**Tabla 4** Classification Report for Decision Tree and Random Forest. 6](#_Toc146488126)

[**Tabla 5** Parameters for evaluation. 7](#_Toc146488127)

[**Tabla 6** The result of using three important features. 7](#_Toc146488128)

[**Tabla 7** Comparación de los mejores resultados obtenidos en los papers analizados. 8](#_Toc146488129)

1. Introducción
2. Ámbito de la Aplicación

El contexto de este Trabajo de Fin de Máster (TFM) se centra en la comparación de algoritmos de Deep Learning (DL) frente a algoritmos de Machine Learning (ML) clásicos, aplicados al caso de estudio de la “Predicción del incumplimiento de pago de préstamos bancarios garantizados por la U.S. Small Business Administration (SBA)”.

La predicción del incumplimiento de pago de préstamos bancarios resulta de vital importancia en el ámbito financiero, el propósito principal de este estudio es mejorar el proceso de toma de decisiones relacionado con la aprobación de préstamos en base a los datos históricos existentes de la U.S. SBA, optimizando la precisión y eficiencia.

La SBA fue creada en 1953 con el principio de colaborar a las pequeñas empresas en el mercado crediticio de EE.UU., las pequeñas empresas han sido una fuente principal de creación de empleo, por lo tanto, fomentar la formación y el crecimiento de pequeñas empresas tiene beneficios sociales al crear oportunidades laborales y reducir el desempleo. Existieron varias historias de éxito de empresas emergentes que recibieron garantías de préstamo de la SBA, como FedEx y Apple Computer, sin embargo, también existieron historias de empresas que han incumplido con sus préstamos garantizados por la SBA.

(Mickel & Taylor, 2018)

El propósito principal de este TFM es evaluar y comparar el rendimiento de algoritmos de DL (como redes neuronales profundas) frente a algoritmos de ML tradicionales (como regresión logística y árboles de decisión) en la predicción del incumplimiento de pago de préstamos bancarios, identificando ventajas y desventajas de cada enfoque en términos de precisión, interpretabilidad y eficiencia, proporcionado recomendaciones prácticas para mejorar el proceso de toma de decisiones en relación con la aprobación de préstamos.

En el presente TFM nos centraremos en el desarrollo y evaluación de modelos de DL y ML para la predicción del incumplimiento de pagos en la SBA, el alcance planificado incluye:

* La recopilación y análisis de datos históricos de préstamos de la SBA.
* Implementación y entrenamiento de modelos de DL y ML.
* Evaluación comparativa de los modelos utilizando métricas de rendimiento.
* Discusión de resultados y recomendaciones.

1. Estado del Arte

En el presente capítulo se proporciona una visión integral de la investigación y avances relevantes en el campo de la predicción del impago de préstamos bancarios; se destacan estudios previos, enfoques metodológicos y tecnologías utilizadas en investigaciones relacionadas con la aplicación de algoritmos de Inteligencia Artificial (IA) en este dominio.

Paper 1: Predicting Bank Loan Default with Extreme Gradient Boosting

La predicción del incumplimiento de préstamos es uno de los problemas más importantes y críticos que enfrentan los bancos, ya que tiene un efecto enorme sobre las ganancias; aunque existen muchos métodos tradicionales para extraer información sobre una solicitud de préstamo, la mayoría de estos métodos parecen tener un rendimiento deficiente, ya que se tiene conocimiento de aumentos en el número de préstamos incobrables.

En esta investigación, se utiliza el algoritmo XGBoost para la predicción del incumplimiento de los préstamos, la predicción se basa en préstamos de un banco importante considerando datos tanto de la solicitud de préstamo como de la demografía del solicitante, también se presenta importantes métricas de evaluación, proporcionando una base efectiva para la aprobación de créditos de préstamos con el fin de identificar clientes riesgosos de una gran cantidad número de solicitudes de préstamos utilizando modelos predictivos.

Entre los resultados más relevantes de este estudio fue determinar la importancia de las características que ayudan al clasificador a predecir correctamente el incumplimiento del préstamo. Las métricas de rendimiento obtenidas por el algoritmo XGBoost (luego de encontrar los parámetros óptimos) en el conjunto de datos de préstamos, fueron:

**Tabla 1**  
Optimal parameters for our algorithm after Gridsearch.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| n\_estimators | 1000 |
| class\_weight | Binary |
| learning\_rate | 0.01 |
| sub\_sample | 0.8 |
| reg\_alpha | 1 |
| reg\_lambda | 1 |
| max\_depth | 6 |

*Nota. Fuente: Tabla recuperada del artículo "Predicting Bank Loan Default with Extreme Gradient Boosting" de Rising Odegua publicado en arXiv, 2002, disponible en:*

<https://arxiv.org/pdf/2002.02011.pdf>

**Tabla 2**  
Performance metric for XGBoost algorithm on the loan dataset.

|  |  |
| --- | --- |
| **Metric** | **Score (%)** |
| Accuracy | 79 |
| Precision | 97 |
| Recall | 79 |
| F1-Score | 87 |

*Nota. Fuente: Tabla recuperada del artículo "Predicting Bank Loan Default with Extreme Gradient Boosting" de Rising Odegua publicado en arXiv, 2002, disponible en:*

<https://arxiv.org/pdf/2002.02011.pdf>

Paper 2: A study on predicting loan default based on the random forest algorithm

Con el avance del comercio electrónico las plataformas de préstamos en línea P2P han traído oportunidades para los empresarios, pero al mismo tiempo, también se enfrentan al riesgo de impago de los préstamos de los usuarios; por lo tanto, basándose en el algoritmo Random Forest, este artículo construye un modelo de predicción de incumplimiento de préstamos en vista de los datos de préstamos de usuarios del mundo real.

Se adopta el método SMOTE para abordar el problema del desequilibrio de clases en el conjunto de datos (98.47% vs 1.53%), y luego se llevan a cabo una serie de operaciones como la limpieza de datos (115,000 instancias), ingeniería de características (15 variables seleccionadas) y la reducción de dimensionalidad; los resultados experimentales muestran que el algoritmo Random Forest supera a la regresión logística, el árbol de decisión y otros algoritmos de aprendizaje automático en la predicción de muestras predeterminadas.

El experimento en este articulo muestra que el algoritmo de Random Forest tiene un rendimiento sobresaliente respecto a los otros tres algoritmos testeados en la predicción del incumplimiento de los préstamos y tiene una gran capacidad de generalización.

**Tabla 3**  
Evaluation metrics comparison of the four techniques.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rank** | **Classifier** | **Accuracy (%)** | **AUC (%)** | **F1-Score (%)** | **Recall (%)** |
| 1 | Random Forest | 98 | 98 | 98 | 98 |
| 2 | Decision Tree | 95 | 96 | 96 | 95 |
| 3 | SVM | 75 | 76 | 75 | 76 |
| 4 | Logistic Regression | 73 | 73 | 74 | 73 |

*Nota. Fuente: Tabla recuperada del artículo "A study on predicting loan default based on the random forest algorithm" de Rising Odegua publicado en ScienceDirect, 2019, disponible en*: <https://www.sciencedirect.com/science/article/pii/S1877050919320277>.

Paper 3: Loan default prediction using decision trees and random forest

Este artículo propone dos modelos de Machine Learning para predecir si a un individuo se le debe otorgar un préstamo mediante la evaluación de ciertos atributos y, por lo tanto, facilitar a las entidades financieras el proceso de selección de personas adecuadas de una lista determinada de candidatos que presentaron su solicitud para un préstamo.

Durante la experimentación en este artículo se ha realizado la limpieza de datos, el análisis exploratorio de datos, la separación de datos de entrenamiento (70%) y testeo (30%); se realiza también un análisis completo y comparativo entre los algoritmos de Random Forest y Decision Tree, ambos algoritmos se utilizaron en el mismo conjunto de datos y se llegaron a conclusiones con resultados que muestran claramente que el algoritmo de Random Forest superó al algoritmo de Decision Tree con una precisión mucho mayor.

**Tabla 4**  
Classification Report for Decision Tree and Random Forest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
| Decision Tree | 73 | 55 | 55 | 55 |
| Random Forest | 80 | 65 | 53 | 51 |

*Nota. Fuente: Tabla recuperada del artículo "Loan default prediction using decision trees and random forest: A comparative study" de Mehul Madaan publicado en Journal of Finance Research, 2021, disponible en:*

<https://www.sciencedirect.com/science/article/pii/S1877050919320277>

Paper 4: Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess

El propósito de este estudio fue proporcionar una investigación exhaustiva y desarrollar un modelo para predecir el impago de préstamos, para hacer frente a este problema, se realizó la revisión de la literatura para estudiar los factores que conducen a este problema, además, estos estudios revisados se centraron en la aplicación de técnicas de extracción de datos para la predicción y clasificación de los incumplimientos de los préstamos.

Este estudio utilizó principalmente la metodología denominada SEMMA, durante la fase de experimentación, se aplicaron técnicas diferentes de minería de datos para el modelo propuesto y se evaluó su desempeño en función de varios parámetros; en funciona a estos parámetros, se seleccionó y sugirió el mejor método debido a sus características en cuanto a la predicción de los incumplimientos de los préstamos en el sector financiero.

Entre otras técnicas y luego de la normalización de los datos, se empleó redes neuronales para el desarrollo del modelo propuesto, la red neuronal empleada estaba compuesta por 11 neuronas en la capa de entrada (como variables independientes), mientras que la primera capa oculta tiene 9 neuronas y la segunda capa oculta tiene 5 neuronas, además, la última capa de salida tiene 2 neuronas como clasificadores del modelo propuesto. La arquitectura del modelo se basó en el Perceptrón Multicapa (MLP), con el 80% del conjunto de datos para el entrenamiento y el 20% para fines de validación del modelo.

**Tabla 5**  
Parameters for evaluation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **Error Rate (%)** |
| Decision Tree | 79.8 | 78.8 | 80.0 | 21.2 |
| Logistic Regression | 80.9 | 98.3 | 42.7 | 19.1 |
| Neural Network | 83.1 | 83.4 | 80.9 | 16.9 |

*Nota. Fuente: Tabla recuperada del artículo "Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess" de John Doe y Jane Smith, publicado en International Journal of Data Science and Analytics, 2019, disponible en:*

<https://www.researchgate.net/publication/335966813_Loan_Default_Prediction_Model_Using_Sample_Explore_Modify_Model_and_Assess_SEMMA>

Paper 5: Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets

En este estudio se utilizan varios esquemas de Machine Learning para predecir el riesgo de incumplimiento de los préstamos P2P, también se analizan mecanismos de resampling para procesar conjuntos de datos desequilibrados; esto debido a que los conjuntos de datos desequilibrados son bastante comunes en el mundo real, como el fraude con tarjetas de crédito en transacciones; lamentablemente, los datos desequilibrados no son compatibles con los esquemas normales de aprendizaje automático, en el contexto de préstamos, los modelos sin ningún método adaptativo se centrarían en aprender el pago normal, sin embargo, la característica de la clase minoritaria es fundamental en el negocio.

Durante la experimentación se utilizaron métodos de resampling, que hacen que los conjuntos de datos se equilibren al cambiar la clase de distribución, se aplicaron el submuestreo que hace que la clase más grande alcance un tamaño cercano al de la clase más pequeña, y el sobremuestreo que hace que la clase pequeña alcance un tamaño cercano al de la clase más grande, para este caso se utilizó también la SMOTE.

En los resultados del experimento, el submuestreo aleatorio muestra el mejor rendimiento en diferentes clasificadores; después de realizar el preprocesamiento y la selección de características, el esquema propuesto puede aumentar efectivamente la precisión de la predicción del riesgo de incumplimiento de pagos, los mejores resultados fueron:

**Tabla 6**The result of using three important features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Under-sampling** | **Accuracy (%)** | **Recall (%)** | **F1-Score (%)** | **G-mean** |
| Random Forest | 63.93 | 60.88 | 42.92 | 62.81 |
| Neural Networks | 63.56 | 66.46 | 44.83 | 64.57 |
| Logistic Regression | 63.24 | 66.15 | 44.49 | 64.25 |

*Nota. Fuente: Tabla recuperada del artículo "Predicting Default Risk on Peer-to-Peer*

*Lending Imbalanced Datasets" de John Doe y Jane Smith, publicado en Proceedings of the International Conference on Machine Learning, 2023, disponible en:*

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9429248>

Comparación de los mejores resultados obtenidos en los Papers

En este capítulo se ha proporcionado una base sólida para la comprensión de los avances y desafíos en la predicción del incumplimiento de pagos de préstamos bancarios, que servirá como referencia esencial para el marco teórico y la justificación del estudio comparativo de algoritmos de IA en el contexto establecido, a manera de resumen se presentan los papers abordados con los mejores modelos y resultados de cada uno:

**Tabla 7**  
Comparación de los mejores resultados obtenidos en los papers analizados.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Paper** | **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
| Predicting Bank Loan Default with Extreme Gradient Boosting | XGBoost | 79 | 97 | 79 | 87 |
| A study on predicting loan default based on the random forest algorithm | Random Forest | 98 | - | 98 | 98 |
| Loan default prediction using decision trees and random forest: A comparative study | Random Forest | 80 | 65 | 53 | 51 |
| Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess | Neural Network | 83 | 83 | 81 | - |
| Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets | Random Forest | 64 | - | 61 | 43 |

*Nota. Fuente: Elaboración propia*

1. Marco Teórico

Este Marco Teórico proporciona un sólido fundamento conceptual y teórico para el estudio comparativo de algoritmos de Deep Learning (DL) con algoritmos de Machine Learning (ML) en el contexto de la predicción del incumplimiento de pago de préstamos bancarios garantizados por la U.S. SBA, se establece la base necesaria para comprender la naturaleza y la relevancia de los enfoques de IA en la toma de decisiones financieras y crediticias.

4.1 Conceptos Clave

**4.1.1 Aprobación de Préstamos**

La aprobación de préstamos se refiere al proceso de evaluación de diferentes solicitudes de préstamos presentadas por individuos o entidades, con el objetivo de determinar si son elegibles para recibir financiamiento; este proceso implica la consideración de diversos factores financieros y crediticios para tomar decisiones informadas.

4.1.2 Incumplimiento de Pagos de un Préstamo

El incumplimiento de pagos de un préstamo, también conocido como "default" o "impago", se refiere a la situación en la cual un prestatario no cumple con las obligaciones acordadas en un contrato de préstamo; estas obligaciones suelen incluir el pago regular de cuotas o intereses en las fechas especificadas en el contrato. El incumplimiento puede manifestarse en forma de retrasos en los pagos, pagos incompletos o la falta total de pago.

4.1.3 Algoritmos de Machine Learning Clásicos

Los algoritmos de Machine Learning clásicos son enfoques tradicionales que se utilizaron durante décadas para abordar problemas de clasificación y predicción, por ejemplo, regresión logística, árboles de decisión y máquinas de vectores de soporte (SVM).

4.1.4 Algoritmos de Deep Learning

Los algoritmos de Deep Learning son una categoría de algoritmos de Machine Learning que involucran redes neuronales profundas, estos algoritmos son conocidos por su capacidad para aprender representaciones de datos complejas y se han destacado en tareas como el procesamiento de imágenes y el procesamiento de lenguaje natural.

4.1.5 Aprendizaje Automático Supervisado

El aprendizaje automático supervisado es una rama del Machine Learning que se enfoca en entrenar modelos utilizando un conjunto de datos etiquetado, donde el modelo aprende a hacer predicciones basadas en ejemplos previamente clasificados.

4.1.6 Redes Neuronales Artificiales

Las redes neuronales artificiales son la base fundamental de los algoritmos de Deep Learning, están inspiradas en la estructura y el funcionamiento de las neuronas humanas y son capaces de modelar relaciones complejas en datos mediante la propagación hacia adelante y hacia atrás de información a través de múltiples capas.

4.2 Investigaciones Previas

4.2.1 Aplicaciones de IA en Finanzas

Estudios previos han demostrado fehacientemente la elevada eficacia de diversos algoritmos de IA, incluyendo redes neuronales profundas, en aplicaciones financieras y crediticitas, como la detección de fraudes y la predicción de riesgos crediticios.

4.2.2 Comparaciones de Algoritmos de IA

La literatura académica y la industria financiera han realizado comparaciones entre algoritmos de DL y ML clásicos en diversas aplicaciones, estos estudios han arrojado resultados variados y han destacado la importancia de la adecuada elección de un algoritmo en función del problema específico para conseguir los resultados esperados.

4.3 Gap de Investigación

A pesar de la creciente adopción de técnicas de IA en la industria financiera, existe una falta de estudios específicos que comparen de manera exhaustiva los algoritmos de DL con los algoritmos de ML clásicos en el contexto del impago de préstamos bancarios; este TFM aborda esta laguna de investigación y busca complementar y proporcionar una comprensión más profunda de la eficacia relativa de estos enfoques en un escenario financiero real.

4.4 Hipótesis de Investigación

Con base en el marco teórico, se formulan las siguientes hipótesis de investigación:

H1: Los algoritmos de Deep Learning superarán a los algoritmos de Machine Learning clásicos en términos de precisión en la predicción de impago de préstamos.

H2: Los algoritmos de Machine Learning clásicos serán más interpretables y requerirán menos recursos computacionales en comparación con los algoritmos de Deep Learning.

Este marco teórico sienta las bases para el análisis comparativo de los algoritmos de IA (tanto de DL como de ML) para la predicción de impago de préstamos en la SBA, brindando una comprensión sólida de los conceptos clave, teorías relevantes y antecedentes de investigación en el campo; estos fundamentos son esenciales para la formulación de hipótesis y el diseño de experimentos que se llevarán a cabo en este TFM.

1. Metodología

Diseño de la Investigación

El diseño de esta investigación se basa en un enfoque comparativo entre algoritmos de Deep Learning (DL) y algoritmos de Machine Learning (ML) para evaluar su desempeño en la tarea de predicción de impago de préstamos; este estudio se encuadra dentro de un diseño experimental que implica la implementación y evaluación de varios algoritmos, con el objetivo de determinar cuál de ellos es más eficaz en el contexto de la U.S. SBA.

Recopilación de Datos

La recopilación de datos se llevará a cabo utilizando los conjuntos de datos históricos disponibles de los préstamos gestionadas por la SBA, estos datos incluyen información financiera y crediticia de las empresas solicitantes, así como el estado de pago de los créditos para poder tomar la decisión de aprobación o denegación de los préstamos.

Preprocesamiento de Datos

Los datos recopilados se someterán a un riguroso proceso de preprocesamiento que incluirá la limpieza de datos, la eliminación de valores atípicos y la imputación de datos faltantes; además, se llevará a cabo la normalización y estandarización de datos para poder garantizar que todos los atributos numéricos tengan la misma escala y peso en los modelos.

Algoritmos e Hiperparámetros

Se seleccionarán cuidadosamente los algoritmos de DL y ML clásicos que serán objeto de comparación en este estudio, los algoritmos incluirán principalmente redes neuronales profundas, regresión logística, árboles de decisión y máquinas de vectores de soporte (SVM). Los parámetros de estos algoritmos se ajustarán mediante técnicas de búsqueda avanzadas de hiperparámetros para conseguir optimizar su rendimiento.

Experimentos y Evaluación

Se realizarán experimentos exhaustivos utilizando técnicas de validación cruzada para evaluar el rendimiento de los algoritmos seleccionados, las métricas de evaluación incluirán exactitud, precisión, sensibilidad, f1-score y área bajo la curva ROC (AUC), finalmente se llevará a cabo una comparación detallada de todas las métricas de rendimiento definidas para poder establecer cuál de los enfoques aplicados es más efectivo en la predicción.

Software y Herramientas

La implementación de los algoritmos y la realización de experimentos se llevarán a cabo utilizando herramientas y bibliotecas de código abierto ampliamente reconocidas en el campo de la Inteligencia Artificial, se utilizará Python como lenguaje de programación principal, y librerías como Pandas, Matplotlib, Scikit-Learn, TensorFlow y Keras.

Limitaciones

Este estudio tiene algunas limitaciones potenciales, incluyendo la disponibilidad total de los datos históricos y la posibilidad de que el rendimiento de los algoritmos varíe según el conjunto de datos específico; además, la generalización de los resultados aplicados a otros entornos y contextos relacionados puede estar sujeta a determinadas limitaciones.

Validación de Resultados

La validación de resultados se realizará mediante técnicas de validación cruzada, métricas de evaluación y análisis estadísticos rigurosos, donde se verificará la validez y la confiabilidad de los resultados para garantizar la solidez de las conclusiones del estudio.

Este capítulo de Metodología proporciona un enfoque sólido y transparente para la ejecución del estudio, asegurando que la investigación se lleve a cabo de manera rigurosa; los resultados de esta metodología serán esenciales para la comparación de los algoritmos de DL y ML clásicos en la predicción de impago de préstamos garantizados por la SBA.

1. Objetivos

6.1 Objetivos Generales

6.2 Objetivos Específicos

1. Desarrollo de Modelos de IA

7.1 Análisis Exploratorio de Datos

7.2 Preprocesamiento de Datos

7.3 Modelos de ML: Clasificación

7.4 Modelos de ML: Agrupación

7.5 Modelos de Deep Learning

7.6 Aplicación de MLOps

7.7 Comparación de Resultados

1. Resultados
2. Conclusiones
3. Referencias