Universidad Pontificia Comillas, ICAI - ICADE

ANÁLISIS COMPARATIVO DE TÉCNICAS DE MACHINE LEARNING EN MODELOS DE RIESGO DE CRÉDITO

Autor: Luis Mielgo Larriba
Director: Roberto Knop Muszynski

MADRID | Junio, 2024

ABSTRACT

This final degree project, entitled "Comparative analysis of Machine Learning techniques in Credit Risk models", examines and compares the effectiveness of traditional scoring models and modern machine learning models in predicting default in SMEs, a crucial area for the stability of financial institutions. With increasing complexity in the financial environment, it is essential to adopt methods that improve the accuracy of credit risk predictions. This study contrasts the performance of advanced machine learning models, such as Logistic Regression and Support Vector Machines, with traditional scoring methods such as the Altman Z-Score, the Elisabetsky Score and the Kanitz Thermometer. The research focuses on the predictive capacity of these models, their interpretability and the relevance of the variables used, thus providing a comprehensive view of their effectiveness and applicability in the current regulatory framework of the financial sector. Using a database of SMEs in Spain, this work adopts a mixed approach that combines quantitative analysis with qualitative evaluations, offering a new perspective on the identification of complex and non-linear patterns in credit behavior.

RESUMEN

Este trabajo de fin de grado, titulado "Análisis comparativo de técnicas de Machine Learning en modelos de Riesgo de Crédito", examina y compara la eficacia de los modelos de scoring tradicionales y modernos modelos de machine learning en la predicción de default en PyMEs, un área crucial para la estabilidad de las instituciones financieras. Con el aumento de la complejidad en el entorno financiero, es fundamental adoptar métodos que mejoren la precisión de las predicciones de riesgo crediticio. Este estudio contrasta el rendimiento de modelos avanzados de machine learning, como la Regresión Logística y las Support Vector Machines, con métodos de scoring tradicionales como el Z-Score de Altman, el Score de Elisabetsky y el Termómetro de Kanitz. La investigación se centra en la capacidad predictiva de estos modelos, su interpretabilidad y la relevancia de las variables utilizadas, proporcionando así una visión integral de su efectividad y aplicabilidad en el actual marco normativo del sector financiero. Utilizando una base de datos de PyMEs en España, este trabajo adopta un enfoque mixto que combina análisis cuantitativo con evaluaciones cualitativas, ofreciendo una nueva perspectiva en la identificación de patrones complejos y no lineales en el comportamiento del crédito.

1. Introducción
    a. Introducción
    b. Pregunta de investigación
    c. Objetivos
    d. Justificación del tema objeto de estudio
    e. Metodología
2. Marco Teórico
    a. Revisión de modelos Machine Learning
    b. Revisión de Modelos de puntuación o Scoring
    c. Análisis de Riesgo de Crédito y Machine Learning
    d. Consideraciones legales y regulatorias
3. Descripción de la Base de Datos
4. Preprocesamiento de datos
5. Generación de los modelos
    a. Regresión Logística
    b. Support Vector Machine
    c. Modelos de Puntuación o Scoring
    d. Obtención de coeficientes del sector de actividad
6. Análisis de los Resultados obtenidos
    a. Resultados de la Regresión Logística
    b. Resultados de la Support Vector Machine
    c. Resultados del Z-Score de Altman
    d. Resultados del Score de Elisabetsky
    e. Resultados del Termómetro de Kanitz
    f. Resultados del impacto de los sectores en la probabilidad de Default
7. Conclusiones
8. Referencias Bibliográficas
9. Anexo
10. Anexo de tablas y figuras

Introducción

En el entorno de la gestión de riesgos bancarios nos encontramos con las siguientes tipologías: riesgo de crédito, riesgo de mercado y riesgo de liquidez. En este estudio nos centraremos en la primera, el riesgo de crédito, puesto que su evaluación precisa se ha convertido en un imperativo para garantizar la estabilidad de las instituciones financieras en un entorno cada vez más complejo. Este Trabajo de Fin de Grado se centra en realizar un análisis comparativo de diversas técnicas de Machine Learning y técnicas de scoring aplicadas en modelos de riesgo de crédito. La adopción de estas técnicas, que van desde la más clásica regresión logística hasta algoritmos avanzados como las redes neuronales, ha marcado un cambio paradigmático en la aproximación a los datos financieros, ofreciendo una nueva perspectiva en la identificación de patrones complejos y no lineales en el comportamiento del crédito.

El objetivo primordial de este trabajo es evaluar y comparar el rendimiento predictivo de los modelos Machine Learning de Regresión Logística y Support Vector Machine con modelos de scoring del Z-Score de Altman, Score de Elisabetsky y Termómetro de Kanitz, además de identificar las variables relevantes y la interpretabilidad y la calidad de los modelos. Se pretende identificar cuáles ofrecen resultados más precisos y confiables en la predicción del riesgo crediticio, así como su relevancia en la práctica financiera actual. Esta evaluación abarca no solo el aspecto técnico de los modelos, sino también su aplicabilidad y conformidad con las normativas vigentes en el sector financiero.

Para alcanzar estos objetivos, utilizaremos una base de datos de pequeñas y medianas empresas en España. La investigación sigue un enfoque mixto que integra el análisis cuantitativo con una evaluación cualitativa de las técnicas.

Pregunta de investigación

¿Cómo se comparan los modelos de machine learning, como la Regresión Logística y Support Vector Machine, con los modelos de scoring tradicionales, como el Z-Score de Altman, el Z-Score de Elisabetsky y el Termómetro de Kanitz, en términos de precisión predictiva, calidad y claridad interpretativa para la evaluación del riesgo de crédito en PyMEs españolas en los años 2008 a 2011?

De esta pregunta de investigación se destilan 3 preguntas más concretas:

- ¿Cuál de los modelos evaluados, machine learning o scoring, demuestra mayor precisión predictiva en la evaluación del riesgo de crédito en PyMEs españolas?
- ¿Qué relevancia e influencia tienen factores como el sector de actividad en la probabilidad de default de las PyMEs españolas durante los años de la recesión y crisis económica de 2008?
- ¿Cómo se comparan los modelos de machine learning y los modelos de scoring tradicionales en términos de interpretabilidad y transparencia en sus procesos de toma de decisiones, y qué impacto tiene esto en su aceptación por parte de reguladores, instituciones financieras y clientes?

Objetivos

Para esta investigación nos definimos como objetivos los siguientes:

- **Evaluar y comparar el rendimiento predictivo de los modelos seleccionados en PyMEs españolas.** En el marco de esta investigación, el objetivo primordial consiste en llevar a cabo una evaluación comparativa meticulosa del rendimiento predictivo de diversos modelos Machine Learning de Regresión Logística y Support Vector Machine con modelos de scoring del Z-Score de Altman, Z-Score de Elisabetsky y Termómetro de Kanitz; para el Análisis del Riesgo de Crédito en distintos escenarios. Este objetivo se orienta no solo a identificar el modelo o los modelos con el mejor rendimiento predictivo de forma aislada, sino también a comprender las razones detrás de este rendimiento y cómo estos modelos pueden ser integrados efectivamente en las prácticas de gestión de riesgo de las instituciones financieras, analizando las particularidades de cada uno de ellos.
- **Describir la influencia del sector de actividad en la probabilidad de default.** Aprovechando información sobre los sectores de actividad de las empresas, sus datos financieros e información sobre sus situaciones de default en los años inmediatamente posteriores al estallido de la crisis de 2008 se puede realizar un análisis de qué sectores de actividad han tenido mayor relevancia e impacto en la probabilidad de default. Esto nos debe llevar a describir aquellos sectores más afectados por la crisis y recesión, que llevó a muchas empresas a la quiebra o al incumplimiento de sus obligaciones de pago.
- **Analizar la interpretabilidad**. La importancia de este objetivo radica en que no sólo es necesario maximizar la precisión predictiva de los modelos, sino también su lógica operativa y la capacidad para explicar de manera transparente y comprensible las decisiones y predicciones que generan. Este objetivo busca garantizar que los modelos no solo sean herramientas predictivas eficaces, sino también transparentes y comprensibles, facilitando así su adopción y confianza por parte de los profesionales del sector financiero, reguladores y clientes, asegurando que las decisiones basadas en estos modelos sean justas, éticas y responsables.

Justificación tema objeto de estudio

En la era actual, marcada por una transformación digital acelerada, las instituciones financieras enfrentan el desafío constante de evaluar el riesgo de crédito de manera eficiente y precisa. Tradicionalmente, la evaluación del riesgo de crédito se ha basado en modelos estadísticos que, aunque útiles, a menudo no capturan la totalidad de la complejidad y dinamismo de los perfiles de riesgo en un entorno económico globalizado y en rápida evolución. En este contexto, las técnicas de machine learning emergen como herramientas poderosas y prometedoras, ofreciendo la capacidad de aprender de grandes volúmenes de datos y adaptarse a patrones cambiantes de comportamiento crediticio.

El propósito de este Trabajo de Fin de Grado es realizar un análisis comparativo de técnicas machine learning como Regresión Logística y Support Vector Machine con modelos de scoring como el Z-Score de Altman, el Z-Score de Elisabetsky y el Termómetro de Kanitz. Este estudio no sólo es relevante por su aplicación práctica en el sector financiero, sino que también contribuye a la literatura académica existente, llenando posibles vacíos sobre la efectividad comparativa de estos métodos avanzados para PyMEs.

También resulta de relevancia conocer aquellos sectores más afectados durante la crisis de 2008. Gracias a los resultados estadísticas derivados de esta investigación y junto con datos recolectados de otras fuentes podremos desarrollar qué sectores tuvieron más propensión a incurrir en default en ese contexto temporal.

Metodología

1. **Selección de la Base de Datos**: Selección de una base de datos relevante y representativa para el estudio de riesgo de crédito, en este caso nuestro interés se centra en información financiera y contable de Pequeñas y Medianas Empresas españolas. La base de datos contiene información financiera, contable y general de las empresas durante los años de la crisis económica de 2008.
2. **Preprocesamiento de Datos**: Realización de un análisis exploratorio de los datos para identificar posibles problemas, como valores atípicos, datos faltantes, redundancias o outliers. Implementaremos técnicas de preprocesamiento, como normalización y binarización de variables categóricas, para garantizar la operabilidad de los datos.
3. **Selección de Variables**: Seleccionaremos aquellas variables más relevantes en la predicción del riesgo crediticio para los modelos de machine learning. Para los modelos de scoring escogeremos las variables que vienen determinadas por la formulación teórica de cada modelo.
4. **Elección de Algoritmos de Machine Learning y Modelos de Puntuación**: Selección de los modelos de Machine Learning, Regresión Logística y Support Vector Machine, y los modelos de puntuación, Z-Score de Altman, Z-Score de Elisabetsky y Termómetro de Kanitz.
5. **División de Datos y Cross Validation**: Dividiremos la base de datos en conjuntos de train y test para evaluar el rendimiento de los modelos. Aplicaremos técnicas de validación cruzada para garantizar la generalización de los resultados, evitando problemas de overfitting.
6. **Entrenamiento y Evaluación de Modelos**: Entrenaremos cada modelo seleccionado utilizando el conjunto de entrenamiento y evaluaremos su rendimiento utilizando métricas adecuadas para la predicción de riesgo crediticio, como precisión, sensibilidad, especificidad, valor predictivo, área bajo la curva ROC y Criterios de Akaike y Criterios Bayesianos.
7. **Análisis Comparativo**: Realizaremos un análisis detallado de los resultados obtenidos, comparando el rendimiento de los diferentes modelos e identificando fortalezas y debilidades de cada técnica en términos de precisión e interpretabilidad.
8. **Interpretación de Modelos**: Analizaremos la precisión, generalización e interpretabilidad de los modelos.
9. **Presentación de Resultados**: Presentaremos los resultados de manera clara y visual mediante gráficos, tablas y métricas de evaluación, destacando las principales conclusiones derivadas del análisis comparativo.
10. **Discusión de Limitaciones y Consideraciones Legales y Regulatorias**: Identificar y discutir posibles limitaciones del estudio, como sesgos en los datos, limitaciones algorítmicas y cuestiones éticas asociadas al uso de modelos de machine learning en el sector financiero.

Marco teórico

Revisión de Modelos Machine Learning

a. Regresión Logística. Es la técnica estadística más utilizada en el Análisis de Riesgo de Crédito (ARC) debido a su capacidad predictiva para la ocurrencia de eventos binarios, como determinar la probabilidad de incumplimiento de un crédito o clasificar a los solicitantes de crédito en categorías de riesgo alto o bajo. En la década de 1980, se adoptaron ampliamente la regresión logística y la programación lineal como técnicas fundamentales para el desarrollo de sistemas de puntuación crediticia. Hoy en día, a nivel global, el campo ha evolucionado para incluir métodos avanzados de inteligencia artificial, tales como sistemas expertos y redes neuronales (Thomas, David, & Crook, Credit scoring and its Applications, 2002). Esta metodología es especialmente efectiva en situaciones donde la variable dependiente es categórica y binaria, por ejemplo, incumplimiento o no incumplimiento. Utiliza una función logística para modelar la probabilidad del evento basándose en variables independientes, que pueden ser tanto cuantitativas (como ingresos o deudas) como cualitativas (por ejemplo, calificación de auditor). La regresión logística trabaja estimando los coeficientes de la combinación lineal de estas variables independientes usando el método de máxima verosimilitud, proporcionando así una estimación de las probabilidades de ocurrencia de un evento específico.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ \dots \ + \beta_n * X_n$$

p: es la probabilidad de que $Y = 1$.

$\frac{p}{1-p}$: es la razón de probabilidades.

$log(\frac{p}{1-p})$: es el logaritmo natural de la razón de probabilidades (logit).

$\beta_0, \beta_1, \beta_2, \dots, \beta_n$: son los coeficientes que se estiman del modelo.

$X_1, X_2, \dots, X_n$: son las variables predictoras independientes.

b. Least Absolute Shrinkage and Selection Operator (LASSO). Es un método de regresión utilizado en el ARC para optimizar la selección de variables y aumentar la precisión predictiva, especialmente útil cuando se manejan grandes cantidades de variables predictoras. Este método aborda problemas comunes como la multicolinealidad y el sobreajuste en grandes conjuntos de datos mediante la imposición de una restricción de penalización L1, que es la suma de los valores absolutos de los coeficientes de regresión. Esta penalización conduce a la reducción de algunos coeficientes a cero, seleccionando efectivamente un subconjunto más relevante y manejable de variables. Lambda es el parámetro que controla la fuerza de la penalización en los métodos de regularización, como L1. Cuanto más baja sea la lambda,

más similar será la función a mínimos cuadrados ordinarios. Cuanto mayor sea la lambda, mayor será el efecto que tenga el término de regularización y menos variables se utilizan en el modelo. Su aplicación es particularmente valiosa en la construcción de modelos de puntuación de crédito, donde se busca un equilibrio entre complejidad, interpretabilidad y precisión predictiva.

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\}$$

n: es el número de observaciones.

y_i: es la variable dependiente.

x_{ij}: es el valor del predictor jjj para la observación iii.

β_0: es el término de intercepción.

β_j: son los coeficientes del modelo para los predictores.

λ: es el parámetro de regularización que controla la cantidad de contracción: cuanto mayor es el valor de λ, mayor es la penalización y por lo tanto mayor es la contracción de los coeficientes hacia cero.

p: es el número de predictores.

c. Random forest: Destaca por su eficacia en análisis complejos y contextos de alta dimensión, como el ARC. Este algoritmo de ensamblaje combina múltiples árboles de decisión, cada uno generado a partir de subconjuntos aleatorios de datos y variables. Esta metodología no solo reduce la varianza y mejora la generalización, sino que también evita el sobreajuste, gracias a la diversidad en la construcción de los árboles. Cada árbol en el bosque aporta su "voto" en el modelo final, lo que permite a Random Forest manejar eficientemente un gran número de variables de entrada y capturar tanto dependencias lineales como no lineales en los datos. Además, esta técnica es notable por su capacidad para identificar las variables más influyentes en la predicción, lo que resulta esencial en contextos como el ARC, donde se requiere discernir entre una amplia gama de factores potencialmente predictivos.

d. Boosting: Es un método de ensamblaje que mejora la precisión predictiva combinando múltiples "weak learners" o modelos simples, como árboles de decisión, que individualmente tendrían un rendimiento justo por encima del azar. Este enfoque se caracteriza por su secuencialidad y corrección de errores, donde cada modelo sucesivo se enfoca en corregir los fallos de sus predecesores, asignando mayor peso a las instancias mal clasificadas previamente. A diferencia del bosque aleatorio, que reduce principalmente la varianza, el boosting es efectivo tanto en la reducción del sesgo como de la varianza, gracias a su proceso iterativo de ajuste y corrección. Incluye un

coeficiente de aprendizaje que controla la influencia de cada modelo en la predicción final, equilibrando el número de iteraciones necesarias.

e. Artificial neural networks (ANN): Inspiradas en la estructura y funcionamiento del cerebro humano, son modelos computacionales avanzados compuestos por neuronas artificiales interconectadas. Estas neuronas procesan y transmiten señales, permitiendo el análisis y modelización de grandes volúmenes de datos, particularmente útiles en el ámbito financiero para el ARC. Caracterizadas por su capacidad de aprendizaje adaptativo, las ANN mejoran su desempeño conforme procesan más información, manejando datos de entrada y generando valores de salida específicos. Su aplicación en el sector financiero destaca por la habilidad de identificar patrones y tendencias complejas en los datos.

f. Support Vector Machines (SVM): Esta técnica construye un hiperplano o conjunto de hiperplanos en un espacio de alta dimensión, que se utiliza para realizar una separación óptima entre las diferentes clases. La particularidad de las SVM es su capacidad para manejar datos no lineales mediante el uso de funciones que transforman el espacio de entrada en un espacio de mayor dimensión donde es más probable que las clases sean linealmente separables. En el ARC, esto permite una segmentación efectiva de los solicitantes de crédito. En este Trabajo utilizaremos un SVM no lineal kernel. En casos de ARC, los datos no siempre son linealmente separables. Para manejar estos casos, las SVM utilizan el "truco del kernel", que permite operar en un espacio dimensional superior donde los datos son linealmente separables sin tener que computar las dimensiones de ese espacio directamente. La función objetivo en este caso es:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

g. Penalized Logistic Tree Regression (PLTR): Esta técnica fusiona LR, que predice la probabilidad de eventos binarios, con un árbol de decisión, que segmenta datos en subgrupos según criterios definidos. Su singularidad radica en la incorporación de una penalización durante la construcción del árbol para contrarrestar el sobreajuste, un fenómeno común en ML donde el modelo se ajusta excesivamente a los datos de entrenamiento, perdiendo capacidad de generalización. Al equilibrar la complejidad del modelo y su habilidad para adaptarse a nuevos datos, facilita una segmentación de clientes basada en el riesgo de manera más precisa y mejora la interpretabilidad y eficacia del modelo en aplicaciones prácticas.

Revisión de Modelos de puntuación o scoring

h. Z-score de Altman. Altman (1968) propuso usar la técnica del análisis discriminante multivariado para la previsión de quiebra de las empresas. El autor estaba interesado específicamente en identificar variables con mayor poder de previsión. El Z-score de Altman es un indicador financiero que predice la probabilidad de quiebra de una empresa. Este modelo fue introducido por Edward I. Altman en 1968 y se basa en una combinación de cinco ratios financieros ponderados que reflejan diversos aspectos de la salud financiera de una empresa, como la liquidez, la rentabilidad, la eficiencia operativa y la estructura del capital. La fórmula del Z-score combina estos ratios en una puntuación única que indica la salud financiera de la empresa. Un Z-score por debajo de 1.8 sugiere un alto riesgo de bancarrota, mientras que un valor por encima de 3 implica solidez financiera . La utilidad del Z-score de Altman radica en su capacidad para proporcionar una medida objetiva y cuantitativa del riesgo de insolvencia de una empresa, lo que lo convierte en una herramienta valiosa tanto para los gestores de empresas como para los inversores y analistas financieros. (Altman, 1968)

$$Z = 1.2 * X_1 + 1.4 * X_2 + 3.3 * X_3 + 0.6 * X_4 + 1.0 * X_5$$

- ➔ $X_1 = \frac{Capital\ de\ Trabajo\ Neto}{Activos\ Totales}$
- ➔ $X_2 = \frac{Reservas\ No\ Distribuidas}{Activos\ Totales}$
- ➔ $X_3 = \frac{Beneficio\ antes\ de\ intereses\ e\ impuestos}{Activos\ Totales}$
- ➔ $X_4 = \frac{Valor\ de\ Mercado\ del\ Capital}{Valor\ Contable\ de\ la\ Deuda\ Total}$
- ➔ $X_5 = \frac{Ventas}{Activos\ Totales}$

$Umbrales$:
$Z > 2.99$: $Zona\ 'segura'$
$1.81 < Z < 2.99$: $Zona\ 'gris'$
$Z < 1.81$: $Zona\ de\ 'peligro'$

i. Termómetro de Kanitz. Kanitz (1978) utilizó la técnica del análisis discriminante y regresión múltiple para proyectar balances futuros. Este modelo se basa en el análisis de diversos indicadores financieros obtenidos de los estados financieros de la empresa, tales como ratios de liquidez, endeudamiento, rentabilidad, y eficiencia operativa. La metodología se enfoca en identificar señales tempranas de dificultades financieras que podrían llevar a una empresa a la insolvencia, permitiendo a los gestores tomar medidas correctivas a tiempo. Se basa en estos índices: Rentabilidad del Patrimonio, Liquidez General, Liquidez Seca, Liquidez Corriente y Grado

de Endeudamiento. Según el criterio de Kanitz, cuando el indicador *Y* resultante de aplicar su fórmula es inferior a -3, sugiere una alarma de posible quiebra para la empresa. Cuanto más negativo sea el número, más crítica es la situación financiera. En contraste, un valor de *Y* por encima de cero indica estabilidad financiera, liberando a la dirección de preocupaciones mayores. Un *Y* que oscile entre 0 y -3 marca una zona de alerta temprana denominada por Kanitz como "penumbra", donde se debe proceder con cuidado. Finalmente, un valor entre 0 y +7 en este indicador señala un estado de solvencia, con el riesgo de quiebra disminuyendo a medida que el valor se incrementa dentro de este margen. (Kanitz, 1976)

$$Y = (0.05 * RP + 1.65 * LG + 3.55 * LS) - (1.06 * LC + 0.33 * GE)$$

- ➔ $Rentabilidad\ del\ Patrimonio = \frac{Beneficio\ Neto}{Fondos\ Propios}$
- ➔ $Liquidez\ General = \frac{Activo\ Total}{Pasivo\ Total}$
- ➔ $Liquidez\ Seca = \frac{Activo\ Corriente - Existencias}{Pasivo\ Corriente}$
- ➔ $Liquidez\ Corriente = \frac{Activo\ Corriente}{Pasivo\ Corriente}$
- ➔ $Grado\ de\ Endeudamiento = \frac{Pasivo\ Corriente + Pasivo\ No\ Corriente}{Fondos\ Propios}$

$Umbrales$:
$Y > 0$: $Zona\ 'segura'$
$(-3) < Y < 0$: $Zona\ 'gris'$
$Y < (-3)$: $Zona\ de\ 'peligro'$

j. Modelo de Ohlson O-Score. Este modelo se basa en la teoría de la valoración de activos y utiliza información contable y de mercado para estimar la probabilidad de quiebra. El O-score de Ohlson incluye varios indicadores financieros: Activos Totales, PIB, Pasivos Totales, Fondo de Maniobra, Pasivos Corrientes, Activos Corrientes, Ingresos netos,... Para el O-Score, cualquier resultado superior a 0,5 sugiere que la empresa incumplirá en dos años. (Ohlson, 1980)

k. Modelo de Elisabetsky. El profesor Roberto Elisabetsky desarrolló en 1976 este modelo enfocado al estudio de operaciones de crédito por parte de los bancos. Utilizó el análisis discriminante para un grupo de 373 empresas del ramo de confecciones, de las que 274 eran empresas en buenas condiciones financieras y 99 presentaban problemas de liquidez. Este modelo se basa en la utilización de 5 variables. Si el resultado es superior a 0,5 la empresa se considera solvente, si es inferior tendrá problemas de solvencia. (Elisabetsky, 1976)

$$SE = 1.93 * X_1 - 0.20 * X_2 + 1.02 * X_3 + 1.33 * X_4 - 1.12 * X_5$$

$$\rightarrow X_1 = \frac{Beneficio\ después\ de\ impuestos}{Ventas}$$

$$\rightarrow X_2 = \frac{Disponible}{Activo\ No\ Corriente}$$

$$\rightarrow X_3 = \frac{Cuentas\ a\ cobrar}{Activo\ Total}$$

$$\rightarrow X_4 = \frac{Existencias}{Activo\ Total}$$

$$\rightarrow X_5 = \frac{Pasivo\ Corriente}{Activo\ Total}$$

$Umbrales$:

$Y > 0.5$: $Zona\ de\ 'solvencia'$

$Y < 0.5$: $Zona\ de\ 'insolvencia'$

En la tabla a continuación se presentan los análisis de distintos modelos machine learning sobre diferentes datasets en base a los datos de 'Area Univer the Curve', 'Power Gain Index', 'Pearson Correlation Coefficient', 'Kolmogorov-Smirnov' y 'Brier Score'.

Methods	AUC	PGI	PCC	KS	BS
Australian dataset					
Linear Logistic Regression	0.8998	0.5664	0.8374	0.7135	0.1186
Non-Linear Logistic Regression	0.6090		0.6067	0.2266	0.3921
Non-Linear Logistic Regression + ALasso	0.8866	0.5092	0.8214	0.6816	0.1333
Random Forest	0.9344	0.6246	0.8603	0.7523	0.0999
PLTR	0.9299	0.6370	0.8606	0.7425	0.1029
Support Vector Machine	0.9210	0.5557	0.8445	0.7391	0.1122
Neural Network	0.9141	0.5799	0.8539	0.7366	0.1102
Taiwan dataset					
Linear Logistic Regression	0.6310	0.2099	0.7586	0.2506	0.2344
Non-Linear Logistic Regression	0.5963	0.0984	0.7035	0.1927	0.2965
Non-Linear Logistic Regression + ALasso	0.7596	0.5029	0.7871	0.3926	0.1447
Random Forest	0.7722	0.4924	0.8102	0.4177	0.1362
PLTR	0.7780	0.5156	0.7959	0.4257	0.1352
Support Vector Machine	0.7102	0.3207	0.8195	0.3382	0.1461
Neural Network	0.7304	0.4226	0.7879	0.3885	0.1401
Housing dataset					
Linear Logistic Regression	0.7904	0.5508	0.8103	0.4450	0.1228
Non-Linear Logistic Regression	0.7965	0.5425	0.8239	0.4650	0.1199
Non-Linear Logistic Regression + ALasso	0.8113	0.5754	0.8217	0.4815	0.1125
Random Forest	0.9387	0.8157	0.9036	0.7455	0.0736
PLTR	0.9011	0.7341	0.8818	0.6694	0.0844
Support Vector Machine	0.7890	0.5514	0.8093	0.4444	0.1254
Neural Network	0.7910	0.5478	0.8132	0.4470	0.1208

Tabla 1

Fuente: *Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds*

Análisis de Riesgo de Crédito y Machine Learning

El riesgo de crédito representa la potencial pérdida que una entidad financiera podría enfrentar debido a que las partes con las que interactúa no cumplan con sus obligaciones contractuales. Las instituciones financieras enfrentan este riesgo al otorgar créditos a sus clientes mediante productos como tarjetas de crédito, préstamos hipotecarios o líneas de crédito.

Comúnmente, los bancos crean modelos para clasificar a sus clientes en distintos niveles de riesgo, que se emplean tanto para establecer límites en los montos de préstamos y créditos, como para aplicar tasas de interés más altas como forma de compensación por el riesgo adicional.

En la gestión de riesgo de crédito, es habitual referirse a conceptos de pérdidas esperadas e inesperadas. La pérdida esperada en una operación se define como el promedio de las posibles pérdidas y se calcula generalmente multiplicando la probabilidad de incumplimiento (probabilidad de que el deudor falle en cumplir sus obligaciones), la exposición al riesgo (cantidad total del crédito otorgado) y la pérdida en caso de incumplimiento (estimación de lo que efectivamente se pierde una vez que se han ejecutado las garantías y otras medidas de recuperación).

El Machine Learning es una rama de la inteligencia artificial cuya metodología implica la creación de modelos algorítmicos que procesan datos históricos para su entrenamiento en función de una necesidad específica y, mediante este proceso, descubren patrones que sirven para anticipar futuros acontecimientos o tendencias. Esto le da a las computadoras la habilidad de “aprender” sin necesidad de ser programadas explícitamente. Según Forbes (2021), el 43% de empresas ha detectado que la implementación de algoritmos ML ha sido más útil de lo que esperaban. Entre las empresas que más implementan estos modelos se encuentran las entidades financieras, ya que el 37% de 60 instituciones financieras tienen modelos de ML completamente operativos dedicados a automatizar los procesos de asignación de crédito, según una encuesta realizada por la EBA en 2020, sobre todo en el más crítico riesgo de crédito según IIF (European Banking Authority, 2021). En este sentido, las instituciones de crédito han orientado su enfoque desde objetivos regulatorios, tales como la estimación del capital requerido, hacia aplicaciones más centradas en sus operaciones comerciales. Esto incluye la toma de decisiones sobre la otorgación de nuevos préstamos, el seguimiento de créditos en curso, la gestión de préstamos con riesgo de impago y la implementación de sistemas de alerta temprana.

Consideraciones legales y regulatorias

La importancia de mantener la calidad de los datos y respetar la privacidad es fundamental en la implementación del Machine Learning en el sector financiero. La Autoridad Bancaria Europea (EBA) señala que la principal restricción del ML es la calidad de los datos, resaltando la preferencia de las entidades financieras por utilizar datos estructurados propios para asegurar el cumplimiento de las normativas de privacidad y la confiabilidad de la información. La capacidad de interpretar los modelos y gestionar los sesgos adquiere una significativa relevancia, influenciando aspectos legales y éticos que afectan la protección de clientes y consumidores, y es considerada principalmente en términos de conducta de mercado. (European Banking Authority, 2023)

El GDPR en su Artículo 22 (Unión Europea, 2016) protege a los individuos de decisiones totalmente automatizadas que puedan afectarles significativamente, exigiendo la inclusión de evaluaciones humanas en procesos clave como la concesión de créditos. La Comisión Europea enfatiza que los resultados de los modelos de ML deben ser comprensibles para todos los involucrados, incluidos los clientes, dada la relevancia económica de las decisiones crediticias en las vidas de las personas. Esto subraya la necesidad de algoritmos explicables que promuevan la transparencia y permitan clarificar las decisiones tomadas. En el ámbito de la gestión de riesgos crediticios, esto implica que las instituciones financieras deben comunicar claramente los factores determinantes en las decisiones de crédito a los clientes.

Controlar los sesgos en los modelos de ML es crucial, como lo define la Comisión Europea, comprometiéndose a asegurar una distribución equitativa de beneficios y costos y a prevenir el sesgo, la discriminación y la estigmatización. Los sesgos pueden originarse de diversas fuentes, como el sesgo de muestra debido al entrenamiento de algoritmos con datos históricos sesgados, o el sesgo por asociación, donde el GDPR limita el uso de datos personales sensibles para evitar discriminaciones. Además, el sesgo algorítmico puede surgir cuando los modelos dan preferencia a ciertas variables sobre otras, lo que podría perjudicar a aquellos sin un amplio historial crediticio pero financieramente solventes.

En febrero de 2024, el BCE emitió una guía destacando la importancia de la gestión adecuada de la calidad de los datos y la protección de la privacidad en el uso de modelos internos dentro del ámbito financiero (BCE, 2024). Esta guía establece que las instituciones financieras deben definir y aplicar estándares rigurosos de calidad de datos, asegurando así la integridad, completitud y relevancia de los datos empleados en dichos modelos. Se enfatiza en varios aspectos de la calidad de los datos, incluyendo su precisión, consistencia, actualidad, unicidad, validez, disponibilidad y rastreabilidad.

La guía propone un sistema de indicadores y controles de calidad que deben ser aplicados de manera sistemática a lo largo de todo el ciclo de vida de los datos, desde su recolección hasta su utilización en informes, abarcando tanto datos históricos como de aplicaciones actuales. Se insta a las instituciones a implementar un proceso estructurado para identificar y corregir cualquier deficiencia en la calidad de los datos, lo cual incluye evaluaciones independientes y la creación de un marco de control efectivo para asegurar la aplicación de procedimientos robustos, especialmente en los procesos manuales.

Este enfoque no solo mejora la precisión y fiabilidad de la información utilizada en decisiones de crédito y gestión de riesgos, sino que también se alinea con las normativas

legales y regulatorias diseñadas para proteger a los consumidores y asegurar una conducta de mercado ética (BCE, 2024).

Descripción de la base de datos

Para realizar este análisis comparativo he optado por utilizar una base de datos de empresas de pequeño y mediano tamaño de España, la cual obtuve de la Universidad Pontificia Comillas para un trabajo de clase de la asignatura de Simulación Financiera para Empresas e Instituciones en el año 2023. La base de datos contiene información de los años 2008, 2009, 2010 y 2011 para 1779 empresas de toda la geografía española. Las 92 variables de las que consta son las siguientes:

- **Información general**: Id, Año, Número empleados, Calificación auditor, Código primario CNAE 2009, Provincia
- **Activo**: Activo no corriente, Inmovilizado intangible, Inmovilizado material, Inversiones inmobiliarias, Inversiones en empresas del grupo y asociadas a largo plazo, Inversiones financieras a largo plazo, Activos por impuesto diferido, Deudas comerciales no corrientes, Activo corriente, Existencias, Deudores comerciales y otras cuentas a cobrar, Clientes por ventas y prestaciones de servicios, Clientes por ventas y prestaciones de servicios a largo plazo, Clientes por ventas y prestaciones de servicios a corto plazo, Accionistas (socios) por desembolsos exigidos, Otros deudores, Inversiones en empresas del grupo y asociadas a corto plazo, Inversiones financieras a corto plazo, Periodificaciones a corto plazo, Efectivo y otros activos líquidos equivalentes, Total activo (A + B)
- **Pasivos y Patrimonio Neto**: Patrimonio neto, Fondos propios, Capital, Capital escriturado, Capital no exigido, Prima de emisión, Reservas, Acciones y participaciones en patrimonio propias, Resultados de ejercicios anteriores, Otras aportaciones de socios, Resultado del ejercicio, Dividendo a cuenta, Ajustes por cambios de valor, Subvenciones, donaciones y legados recibidos, Pasivo no corriente, Provisiones a largo plazo, Deudas a largo plazo, Deudas con entidades de crédito, Acreedores por arrendamiento financiero, Otras deudas a largo plazo, Deudas con empresas del grupo y asociadas a largo plazo, Pasivos por impuesto diferido, Periodificaciones a largo plazo, Acreedores comerciales no corrientes, Deuda con características especiales a largo plazo, Pasivo corriente, Provisiones a corto plazo, Deudas a corto plazo, Deudas con entidades de crédito, Acreedores por arrendamiento financiero, Otras deudas a corto plazo, Deudas con empresas del grupo y asociadas a corto plazo, Acreedores comerciales y otras cuentas a pagar, Proveedores, Proveedores a largo plazo, Proveedores a corto plazo, Otros acreedores, Periodificaciones a corto plazo, Deuda con características especiales a corto plazo, Total patrimonio neto y pasivo (A + B + C)
- **Cuenta de Resultados:** Importe neto de la cifra de negocios, Variación de existencias de productos terminados y en curso de fabricación, Trabajos realizados por la empresa para su activo, Aprovisionamientos, Otros ingresos de explotación, Gastos de personal, Otros gastos de explotación, Amortización del inmovilizado, Imputación de subvenciones de inmovilizado no financiero y otras, Excesos de provisiones, Deterioro y resultado por enajenaciones del inmovilizado, Otros resultados, Resultado de explotación (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12), Ingresos financieros, Imputación de subvenciones, donaciones y legados de carácter financiero,

Otros ingresos financieros, Gastos financieros, Variación de valor razonable en instrumentos financieros, Diferencias de cambio, Deterioro y resultado por enajenaciones de instrumentos financieros, Resultado financiero (13 + 14 + 15 + 16 + 17), Resultado antes de impuestos (A + B), Impuestos sobre beneficios, Resultado del ejercicio (C + 18),

- **Variable Objetivo**: Default

Como podemos ver en el siguiente gráfico de barras con la frecuencia del número de empleados, la mayoría de empresas de la base de datos son PyMEs, y las empresas que tienen un número de empleados mucho mayor que las demás serán eliminadas en la fase de preprocesamiento.

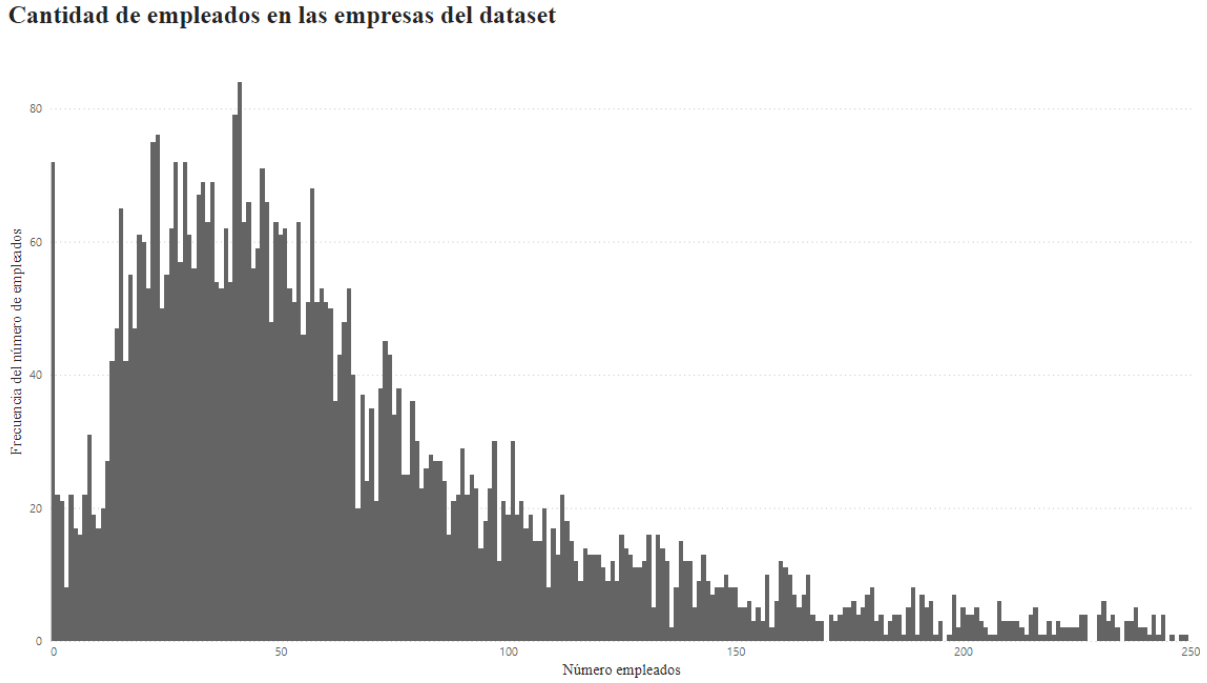

Figura 1

El conjunto de datos tiene 92 columnas y 7117 filas, esto es porque cada fila corresponde a los datos de un año de cada empresa, por lo cual a cada empresa le corresponden 4 filas, sirviendo la columna Id como identificador y la columna Año como el año correspondiente (2008, 2009, 2010 o 2011).

No se debe pasar por alto que estos son los años en los cuales se produjo la crisis económica mundial producida por la explosión de la burbuja inmobiliaria, y subsecuente crisis bancaria. Esto supuso una congelación en el crédito bancario y descenso drástico del valor del mercado inmobiliario, lo que hizo que muchas empresas quebrarasen, aumentando el desempleo y empobreciendo a la población.

El Producto Interior Bruto (PIB) de España experimentó una caída sostenida en la segunda mitad de 2008, lo que llevó al país a entrar en una recesión después de quince años de crecimiento económico continuo. Aunque España logró salir brevemente de la recesión en el segundo trimestre de 2010, el PIB comenzó a disminuir nuevamente a partir de 2011. Como resultado de estas recesiones consecutivas, el PIB per cápita de España, que en 2006

superaba en un 5% la media de la Unión Europea, descendió hasta representar solo el 95% de esta media en 2013. (El Mundo, 2009)
Desde su introducción en 1962, el ratio de morosidad en España alcanzó niveles récord, superiores incluso a los de la crisis económica de 1993. En diciembre de 2013, este ratio llegó a su máximo histórico con un 13,62 %. A pesar de que a finales de 2012 el traspaso de activos tóxicos a la Sareb logró reducirlo en casi un punto porcentual, en enero de 2014 el Banco de España implementó un cambio en la metodología de cálculo de este ratio, excluyendo a los Establecimientos Financieros de Crédito de las entidades de crédito, lo que provocó una ligera caída en febrero de ese año. (El Mundo, 2014)

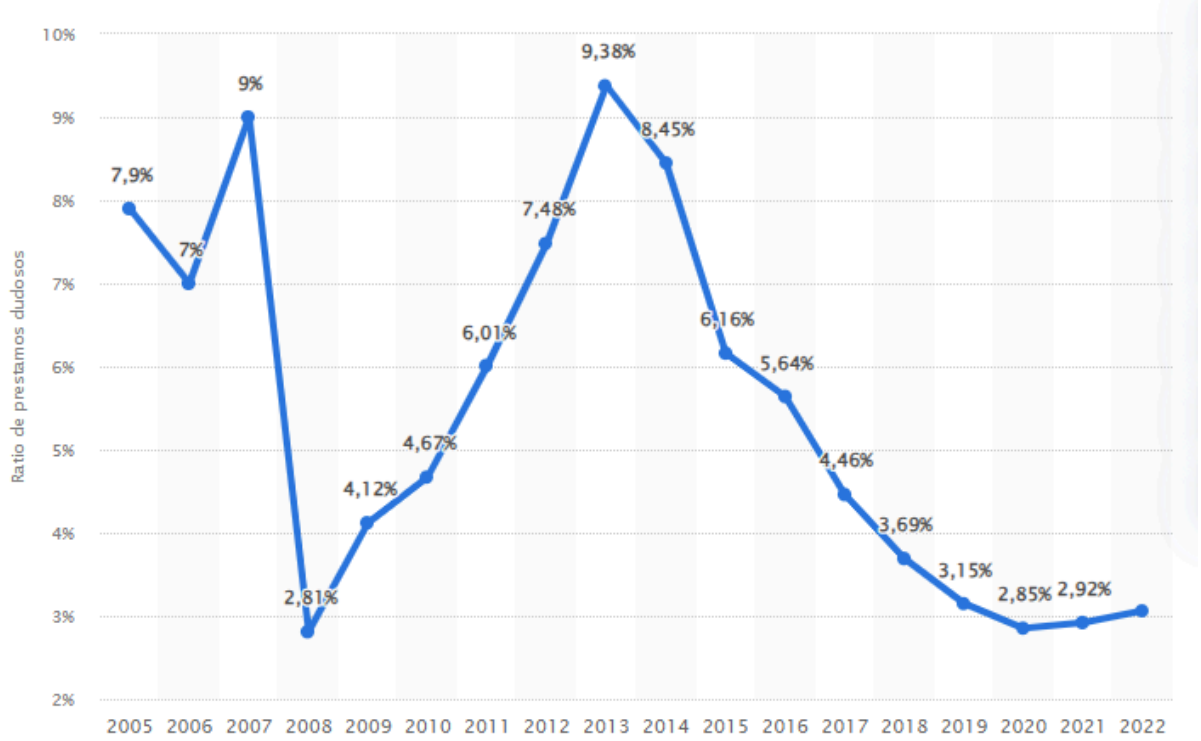

Figura 2

Fuente: Statista, 2024

Al trabajar solo con datos de crisis, los modelos podrían estar sobreajustados a las condiciones extremas, lo que podría no ser representativo de los ciclos económicos normales. Esto puede llevar a predicciones sesgadas cuando se aplican en tiempos de estabilidad económica. Durante una crisis, ciertos indicadores financieros como los ratios de endeudamiento pueden comportarse de manera atípica. Los modelos pueden interpretar incorrectamente estos indicadores como signos de alto riesgo, aún cuando las empresas puedan ser fundamentalmente sólidas.
Los modelos pueden carecer de generalización al ser aplicados en escenarios fuera de una crisis debido a su entrenamiento en un conjunto de datos muy específico. Sin factores de mercado, los modelos podrían perderse señales cruciales sobre las condiciones económicas generales que afectan a todas las empresas, como las tasas de interés, inflación, y cambios en la legislación.

La columna de Calificación Auditor tiene cuatro categorías: Aprobado, Salvedades, Desfavorable y Denegado. Estas categorías de auditoría provienen de normativas de auditoría financiera establecidas por organismos profesionales de contabilidad y auditoría a nivel nacional e internacional. Estas categorías reflejan la opinión del auditor sobre los estados financieros de la entidad auditada, indicando si estos representan de manera fiel la situación financiera de la empresa, su rendimiento y sus flujos de efectivo.

- Aprobado (sin salvedades): También conocido como "opinión limpia", indica que los estados financieros presentan de manera justa y adecuada la posición financiera y los resultados de las operaciones de la empresa, en conformidad con los principios de contabilidad generalmente aceptados.
- Salvedades: Indica que, aunque los estados financieros son una representación justa en su mayoría, hay ciertos aspectos que no cumplen completamente con los principios de contabilidad generalmente aceptados o que no han podido ser verificados completamente.
- Desfavorable (opinión adversa): El auditor cree que los estados financieros no presentan de forma justa la posición financiera, los resultados de las operaciones o los flujos de efectivo de la empresa, en conformidad con los principios de contabilidad generalmente aceptados. Esta opinión se da cuando las distorsiones son tan significativas que impiden la presentación justa.
- Denegado (dictamen con opinión denegada o abstención de opinión): El auditor no puede obtener suficiente evidencia de auditoría para fundamentar una opinión sobre los estados financieros. Esto podría ser debido a limitaciones significativas del alcance en la auditoría, como la incapacidad del auditor para confirmar ciertas cuentas o transacciones.

Estas categorías vienen detalladas en las Normas Internacionales de Auditoría (NIA), concretamente en la NIA 705 (ICAC, 2013). Estas normas son emitidas por la Federación Internacional de Contadores (IFAC) y son ampliamente adoptadas o adaptadas por países de todo el mundo.

La columna de código primario CNAE 2009 se refiere a la Clasificación Nacional de Actividades Económicas 2009, que es un sistema utilizado en España para clasificar las actividades económicas de las empresas y otros tipos de organizaciones económicas. Este sistema es parte de una estructura armonizada a nivel europeo, similar al NACE (Nomenclatura de Actividades Económicas de la Comunidad Europea), que facilita comparaciones económicas y estadísticas tanto a nivel nacional como europeo. (INE, 2009). Gracias a estos datos del CNAE podremos realizar el análisis sectorial de la influencia de los sectores económicos en la probabilidad de default durante este periodo temporal, y por tanto obtener su relevancia en los defaults ocurridos en los años inmediatamente posteriores a la crisis.

Los sectores correspondientes al Sector primario son aquellos cuyos dos primeros dígitos van del 01 al 09, los correspondientes al Sector secundario del 10 al 35 y del 41 al 43, los correspondientes al Sector terciario del 36 al 39, del 45 al 56, del 64 al 68, del 77 al 82 y del 86 al 88; los correspondientes al Sector cuaternario del 58 al 63, del 69 al 75 y el 85 y los restantes corresponden al Sector quinario.

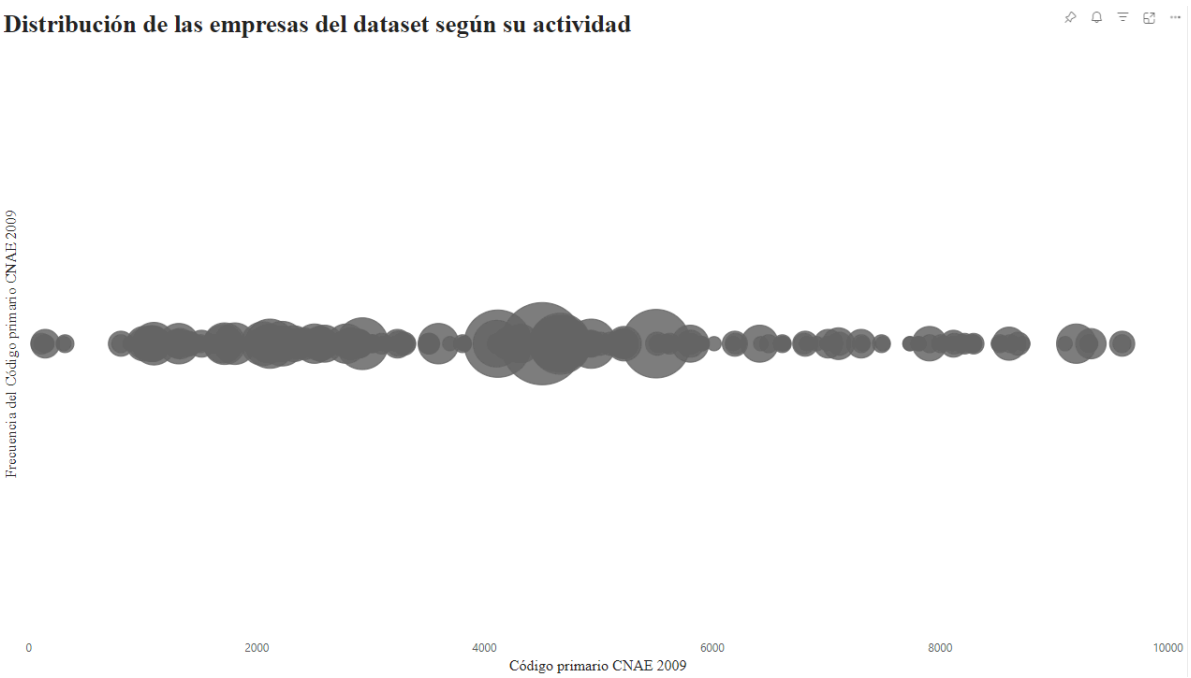

Figura 4

En la columna de provincias viene el nombre de la provincia correspondiente a cada empresa.

La última columna, Default, es una columna binaria en la cual '1' implica que la empresa ha hecho incumplido en sus obligaciones de pago de intereses y/o principal en el plazo de Febrero de 2012 hasta Abril de 2013, es decir, que la empresa haya incurrido en 'Default' en los 15 meses posteriores a los datos que trabajamos. '0' implica que la empresa ha estado al corriente de sus obligaciones de pago durante el mismo periodo.
Ésta columna Default será nuestra variable objetivo para entrenar los modelos de Regresión Logística y Support Vector Machine, es decir, será la variable dependiente que se pretende predecir en el modelo analítico.

Preprocesamiento de datos

En España, la clasificación de una empresa como pequeña y mediana empresa (PyME) se basa en criterios específicos que están alineados con las definiciones proporcionadas por la Unión Europea. Según la recomendación de la Comisión Europea 2003/361/EC, una PyME se define en función de tres criterios principales: número de empleados, facturación anual y/o balance general anual. (Comisión Europea, 2003)

- Microempresa: Menos de 10 empleados y un volumen de negocio o balance general anual que no supere los 2 millones de euros.
- Pequeña empresa: Menos de 50 empleados y un volumen de negocio anual que no exceda los 10 millones de euros o un balance general anual que no exceda los 10 millones de euros.
- Mediana empresa: Menos de 250 empleados y un volumen de negocio anual que no exceda los 50 millones de euros o un balance general anual que no exceda los 43 millones de euros.

Es por eso que para asegurarnos que solo trabajamos con las PyMEs del dataset, no utilizaremos aquellas empresas cuyo Número de Empleados sea mayor a 250 y que tengan un Importe Neto de la Cifra de Negocios inferior a 50 millones de euros o un Total de Activo inferior a 43 millones de euros.

Tanto para el preprocesamiento como para el procesamiento de los datos he utilizado Python con una serie de librerías que detallo en los anexos de este trabajo.
Mediante el uso de la librería Pandas se lee el Excel en el que está la base de datos, sustituyendo los valores vacíos por ceros y simplificando el nombre de las variables.
Después se seleccionaron las variables específicas a utilizar en el modelo de Regresión Logística y en el Support Vector Machine, ya que el uso de demasiadas variables implica:

- Riesgo de Sobreajuste: Uno de los problemas más comunes cuando se utilizan muchas variables, especialmente si la cantidad de datos de entrenamiento no es suficientemente grande en comparación con el número de variables. Esto ocurre cuando el modelo aprende no solo las relaciones subyacentes sino también el ruido en los datos de entrenamiento, lo cual afecta negativamente su capacidad para generalizar a nuevos datos.
- Costo Computacional: El Support Vector Machine, especialmente con un gran número de características, puede ser computacionalmente costoso. Su entrenamiento implica la optimización de un problema cuadrático, que puede volverse más desafiante a medida que aumenta el número de variables.
- Dificultad en la interpretación: A medida que aumenta el número de variables, puede ser más difícil entender cómo cada variable contribuye a las decisiones del modelo.
- La Maldición de la Dimensionalidad: Este es un fenómeno donde al aumentar el número de dimensiones (variables), el volumen del espacio aumenta tan rápidamente que los datos disponibles se vuelven escasos. Esto es problemático porque los modelos de aprendizaje automático dependen de la densidad de los datos para hacer inferencias adecuadas.

Por lo tanto, a coste de perder desempeño predictivo y capturar una menor complejidad, se decide seleccionar un número limitado de variables, quedándonos en este caso con 28.
El criterio de selección de variables se fundamenta en una evaluación preliminar de su importancia. Se han priorizado categorías generales de cuentas financieras de activos, pasivos y resultados. Esta elección se basa en la relevancia aparente que estas categorías poseen dentro del contexto de análisis de riesgo, descartando así el uso de cuentas específicas en favor de un enfoque más holístico y representativo de la situación financiera global de una empresa.
Las 28 variables escogidas son: Número empleados, Calificación auditor, Código primario CNAE 2009, Provincia, Activo no corriente, Activo corriente, Patrimonio neto, Fondos propios, Reservas, Pasivo no corriente, Deudas con entidades de crédito, Acreedores comerciales no corrientes, Pasivo corriente, Provisiones a corto plazo, Deudas a corto plazo, Deudas con entidades de crédito, Acreedores comerciales y otras cuentas a pagar, Importe neto de la cifra de negocios, Otros ingresos de explotación, Gastos de personal, Amortización del inmovilizado, Resultado de explotación, Diferencias de cambio, Resultado financiero, Resultado del ejercicio. Todas son de tipo numérico excepto Calificación auditor, Código primario CNAE 2009 y Provincia.
Las 25 variables numéricas fueron normalizadas, ya que algunas variables tienen rangos de valores mucho mayores que otras debido a la diferencia en tamaños entre empresas, haciendo que los métodos de estimación y optimización sean ineficaces y sesgados hacia las variables con rangos más amplios. Además, puesto que hemos optado por aplicar una penalización Lasso al modelo, las variables han de estar normalizadas. Lasso es una técnica de regresión que incluye un término de penalización en función del coste del modelo. Este término de penalización es la suma de los valores absolutos de los coeficientes multiplicada por un parámetro de regularización, denotado como lambda. Para que Lasso funcione de manera eficaz, las variables deben ser normalizadas para que cada una tenga media cero y desviación estándar uno. Esto asegura que la penalización lambda se aplique de manera equitativa a todas las variables, reflejando mejor la importancia relativa de cada una en el modelo sin ser influenciada por diferencias en la escala.
Las 3 variables categóricas fueron binarizadas. La regresión logística, como muchos otros modelos estadísticos y de machine learning, requiere que las entradas sean numéricas. Al binarizar las variables categóricas, cada categoría se convierte en una nueva variable o característica binaria que indica la presencia (1) o ausencia (0) de esa categoría en cada observación. Esto permite que el modelo evalúe la influencia específica de cada categoría de forma independiente en la predicción del resultado. Si las categorías se codifican de forma ordinal (como 1, 2, 3, etc.), el modelo podría interpretar erróneamente que estas etiquetas tienen un orden o una jerarquía, lo que podría llevar a resultados engañosos.

Para los modelos de scoring se ha hecho el mismo proceso para los datos, menos el paso de normalizar las variables numéricas. Puesto que los modelos que trabajamos tienen umbrales específicos determinados y sus parámetros son ratios sobre las variables, normalizar no ayudaría en nada al proceso. Además que las variables que utilizaremos ya comparten unidades comunes.

Generación de los modelos

Regresión Logística

Empezamos generando el modelo de Regresión Logística. En el dataset hay una desproporción significativa entre las clases del target, donde la clase 0 (sin default) está sobrerrepresentada y la clase 1 (default) infrarrepresentada. En las aplicaciones financieras de predicción de default, los eventos de default son relativamente raros comparados con los casos de no default. Si un modelo se entrena con datos no balanceados, es probable que se sesgue hacia la clase mayoritaria (en este caso, los no defaults), haciéndolo muy bueno en predecir la clase mayoritaria pero pobre en detectar la clase minoritaria, que es en este caso la de mayor interés. Balancear las clases ayuda a asegurar que el modelo no solo sea específico (identificar correctamente los no defaults) sino también sensible (capaz de detectar los defaults). Sin balance, el modelo podría tender a ignorar sistemáticamente la clase minoritaria debido a su escasa representación en el conjunto de datos.
Para subsanar esta situación optamos por proceder al submuestreo de la clase mayoritaria, que consiste en reducir el número de observaciones de las empresas sin Defaults para equilibrar la proporción entre clases. Decidimos reducir la clase mayoritaria para igualar el número de muestras en la clase minoritaria y formar un conjunto equilibrado.

Una vez balanceado el conjunto de datos, dividimos la muestra entre un conjunto de entrenamiento (train set) y un conjunto de prueba (test set). El objetivo principal de dividir los datos en conjuntos de entrenamiento y de prueba es evaluar la capacidad del modelo para generalizar a nuevos datos que no se utilizaron durante el entrenamiento. Esto es crucial porque un modelo que simplemente memoriza los datos de entrenamiento (sobreajuste) podría no funcionar bien en datos no vistos, lo que limita su utilidad práctica.
Generalmente, los datos se dividen aleatoriamente en un 80% de conjunto de entrenamiento y un 20% de conjunto de prueba, por lo que decidimos hacerlo así.

Para mejorar la generalización del modelo y evitar el sobreajuste cuando el número de variables predictoras es alto en relación al número de observaciones o cuando las variables son colineales, se puede aplicar una penalización al modelo de regresión. La penalización consiste en añadir un término adicional a la función de pérdida (o coste) durante el proceso de entrenamiento, que influye en la estimación de los coeficientes del modelo. En un modelo de regresión logística sin penalización, se busca estimar los parámetros que minimicen la función de coste, que típicamente es la log-verosimilitud negativa de las observaciones bajo el modelo.
Existen dos tipos de penalizaciones: Lasso (Least Absolute Shrinkage and Selection Operator) (L1) y Ridge (L2). La penalización Lasso añade el valor absoluto de la magnitud de los coeficientes como término de penalización mientras que la penalización Ridge añade el cuadrado de la magnitud de los coeficientes como término de penalización a la función de coste. Ambos tipos de penalización tienden a reducir la magnitud de los coeficientes, lo cual puede ayudar a reducir el sobreajuste y mejorar la capacidad de generalización del modelo.
Aunque ya hayamos hecho una considerable reducción de variables, puede ser que algunas

de las cuales hayamos escogido no sean relevantes para predecir el default, especialmente al binarizar las variables categóricas hemos generado muchas variables de contenido binario que pueden ser irrelevantes. Lasso ayuda a identificar las variables más influyentes al reducir los coeficientes de las variables menos importantes a cero. Esto efectivamente elimina estas variables del modelo, simplificando el modelo final y haciendo que sea más fácil de interpretar y gestionar.
Otro aspecto a tener en cuenta de la selección de variables que hemos realizado es que muchas están considerablemente correlacionadas. Por ejemplo, cuanto mayor es la variable Inmovilizado Intangible mayor es la cuenta de Activo No Corriente, o cuanto mayor es el Importe Neto de la Cifra de negocios mayor es el Resultado del Ejercicio, generalmente. Ya que el conjunto de datos contiene variables que podemos llamar redundantes, Lasso puede identificar y eliminar estas por medio de la reducción de sus coeficientes a cero, algo que Ridge no hace directamente. Aunque tanto Lasso como Ridge pueden manejar la multicolinealidad, Lasso lo hace de una manera que facilita la selección de variables, lo cual es menos directo en el caso de Ridge que tiende a reducir todos los coeficientes proporcionalmente sin llegar a cero. Al reducir la cantidad de variables en el modelo, Lasso ayuda a prevenir el sobreajuste, haciendo un modelo más simple que generalmente generaliza mejor a nuevos datos no vistos.
En resumen, elegimos Lasso sobre las otras opciones al valorar la capacidad de reducir la complejidad del modelo eliminando activamente las variables no contribuyentes, al buscar claridad en la interpretación de los resultados, y para construir un modelo que sea eficiente contra el sobreajuste en el contexto de la predicción del default financiero.

Una vez seleccionada la penalización Lasso seleccionamos el solucionador ‘liblinear’ de la librería ‘scikit-learn’ de Python. Este solver es utilizado para conjuntos de datos pequeños y medianos y soporta la penalización L1, además de ser el adecuado para trabajar un problema de clasificación binaria como este.

Para ejecutar el modelo utilizamos la función ‘LogisticRegression’ de la librería ‘scikit-learn’, lo que nos permite ajustar el modelo de regresión logística a los datos. Este modelo es capaz de predecir la probabilidad de que las observaciones pertenezcan a una de las clases basándose en los predictores proporcionados y habiendo seleccionado ‘liblinear’, ‘penalty= l1’ y ‘max_iter=10000’, asegurando que el algoritmo tenga suficientes iteraciones para converger.

Para medir la calidad del modelo calculamos AIC (Criterio de Información de Akaike), BIC (Criterio de Información Bayesiano) y Log-Loss (Pérdida Logarítmica).
El Criterio de Información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico para un conjunto de datos. Penaliza la complejidad del modelo (el número de parámetros) junto con el ajuste del modelo (log-likelihood). Un valor bajo AIC generalmente indica un mejor modelo, sugiriendo que es suficientemente complejo para capturar la estructura subyacente de los datos pero no tanto como para sobreajustarse. (Akaike, 1974)
Similar al AIC, el BIC también mide la calidad del modelo, pero con una penalización más

fuerte por el número de parámetros. Está basado en principios bayesianos. Un BIC más bajo indica un modelo que, de manera similar, equilibra bien la complejidad y el ajuste, pero con un enfoque más fuerte en evitar el sobreajuste. (Schwarz, 1978)
En cuanto al Log-Loss, un valor más bajo también indica un mejor modelo. A diferencia del AIC y BIC, que son útiles para la selección de modelos basados en la probabilidad de los datos bajo el modelo, el Log-Loss mide directamente qué tan bien el modelo predice la clase correcta, proporcionando una evaluación de la probabilidad predicha frente a la observación actual, penalizando las predicciones incorrectas. Un Log-Loss bajo significa que el modelo asigna altas probabilidades a las clases correctas y está más seguro de sus predicciones.

Después de procesar el modelo, elaboramos una curva ROC (Receiver Operating Characteristic) y calculamos el AUC (Área Bajo la Curva). La Curva ROC es un gráfico que representa la relación entre la tasa de verdaderos positivos (TPR, o sensibilidad) y la tasa de falsos positivos (FPR, 1 - especificidad) de un clasificador binario. Se traza variando el umbral de decisión que determina cómo se clasifican las probabilidades en clases.
El AUC mide la capacidad del modelo para discriminar entre las clases positivas y negativas. Un AUC de 1.0 representa un modelo perfecto que clasifica correctamente todos los positivos y negativos. Un AUC de 0.5 indica un rendimiento no mejor que el azar. Cuanto mayor sea el AUC, mejor será el modelo en predecir ceros como ceros y unos como unos.
Ambos proporcionan una medida clara de cómo se desempeña el modelo en todas las clasificaciones de umbral y facilitan la comparación entre diferentes modelos o configuraciones del modelo sobre la misma tarea de clasificación, identificando cual tiene un mejor rendimiento.

Además de la curva ROC y el AUC, generaremos una matriz de confusión. Una matriz de confusión es una tabla que permite la visualización del rendimiento de un algoritmo de clasificación. Las filas representan las clases reales, y las columnas las clases predichas. Los elementos de la matriz son TP (verdaderos positivos), FN (falsos negativos), FP (falsos positivos), y TN (verdaderos negativos). En este caso hemos optado por visualizar la proporción en porcentaje correspondiente a cada una de esas cuatro categorías, no su valor absoluto.
Esto nos permite analizar el número de aciertos y errores en cada clase, mostrando cómo el modelo es capaz de manejar cada una de ellas. Además. facilita el cálculo de métricas como la precisión, la sensibilidad, la especificidad y el valor predictivo. Analizando la acumulación de errores o aciertos en cada una de las clases, podemos identificar si el modelo tiene sesgos hacia alguna clase, como clasificar la mayoría de los casos como la clase mayoritaria ignorando la minoritaria o viceversa.

Support Vector Machine

Una vez elaborado el modelo de regresión logística elaboramos la Support Vector Machine o SVM. La máquina de vectores de soporte (SVM) es un modelo predictivo que busca encontrar un hiperplano en un espacio de múltiples dimensiones que mejor separe dos clases de datos. En el caso que nos encontramos se expone en su forma más simple, destinada a la clasificación binaria, en la que el objetivo de la SVM es construir un hiperplano que maximice el margen entre los puntos de datos más cercanos de ambas clases, que son conocidos como vectores de soporte. Este modelo se basa en el concepto de márgenes de decisión, donde el margen se define como la distancia entre el hiperplano de separación y los puntos más cercanos de cada clase. Idealmente, un mayor margen ofrece una mejor generalización, lo que significa una menor probabilidad de error en la clasificación de nuevos datos. Para entrenar este modelo vamos a utilizar las mismas variables normalizadas y balanceadas que usamos para el modelo de regresión. Esto es debido a que la SVM tiene una complejidad computacional que puede ser bastante alta, especialmente para grandes conjuntos de datos con muchas variables, además de para evitar sobreajuste y la maldición de la dimensionalidad.

La capacidad de control de complejidad del modelo está incorporada en SVM mediante el parámetro de regularización *C*, que balancea el error de clasificación en el entrenamiento contra la maximización del margen de decisión. Un *C* grande puede llevar a un modelo de baja sesgo y alta varianza (sobreajuste), mientras que un *C* pequeño puede conducir a un modelo de alta sesgo y baja varianza (sobreajuste). Para ello optimizamos este parámetro utilizando la validación cruzada mediante un enfoque de búsqueda secuencial.
La validación cruzada implica dividir el conjunto de datos completo en *k* subconjuntos, en este caso 5, de aproximadamente igual tamaño. El modelo se entrena 5 veces. Cada vez, uno de los 5 subconjuntos se utiliza como conjunto de prueba (para validar el modelo), y los 5-1 subconjuntos restantes se utilizan como conjunto de entrenamiento. De esta manera, cada subconjunto se utiliza exactamente una vez como conjunto de prueba. Después de entrenar el modelo en cada uno de los pliegues, se obtienen los resultados de rendimiento del modelo, que generalmente son errores de validación calculados en cada pliegue de prueba. Estos errores son promediados para obtener una medida única de rendimiento.

Para la construcción del modelo determinamos que el tipo de kernel a usar es RBF, que es efectivo para espacios de características no lineales. El kernel RBF puede manejar casos donde la relación entre las etiquetas de clase y los atributos es compleja. Además hacemos que el modelo estime las probabilidades, utilizando un parámetro propio de la fórmula que, aunque hace que el modelo sea un poco más lento debido a la necesidad de calibrar las probabilidades internamente usando validación cruzada, calcula y almacena las probabilidades para poder obtenerlas después del entrenamiento.

Una vez obtenidas estas probabilidades hacemos la matriz de confusión y dibujamos la curva ROC con el valor del AUC.

Modelos de Puntuación o Scoring

Para los Modelos de Puntuación o Scoring se recuperan los valores de las variables puras, sin la normalización ni el balanceo de clases. Una vez obtenidas las variables se construyen las fórmulas correspondientes a cada modelo en el código de Python, que son las desglosadas en el apartado de Revisión de Modelos de Puntuación o Scoring en el Marco Teórico anteriormente expuesto.
Se incluye en umbral que los autores determinaron para cada modelo y se procede a la clasificación de los datos en 1: Default o 0: No Default.
Una vez obtenidos los resultados se genera la matriz de confusión y la curva ROC con su AUC.

Obtención de coeficientes del sector de actividad

En un modelo de regresión logística, los coeficientes (β) representan la relación entre cada predictor independiente y el logaritmo de las probabilidades (log-odds) de la variable dependiente (en este caso, la probabilidad de default). Los coeficientes indican cuánto cambia el log-odds del resultado (default) por una unidad de cambio en el predictor, manteniendo todos los demás predictores constantes.
Un coeficiente positivo aumenta el log-odds del default, lo que indica que un aumento en la característica sociedad incrementa la probabilidad de default, mientras que un coeficiente negativo indica lo contrario. Un coeficiente cero implica que la característica no tendría efecto sobre la probabilidad de default.
Cuanto mayor es el valor absoluto del coeficiente, mayor es el impacto que la característica tiene sobre la probabilidad de default en el modelo, ya sea positivo o negativo.

Para obtener estos coeficientes primero escogimos cinco sectores de actividad para englobar los cien determinados por el código CNAE. Estos son: Sector primario (Agricultura, Ganadería, Pesca, Silvicultura, Minería), Sector secundario (Industria Manufacturera, Construcción, Industria, Química, Industria Metalúrgica, Industria Automotriz), Sector terciario (Comercio, Transporte y Logística, Hostelería y Turismo, Educación,Salud, Finanzas y Seguros,Telecomunicaciones, Tecnologías de la Información), Sector cuaternario (Investigación y Desarrollo (I+D), Tecnología de la Información y Comunicación (TIC),Consultoría y Asesoría, Educación) y Sector quinario (Gobierno y Administración Pública,ONGs y Organizaciones Internacionales, Servicios de Alta Dirección, Cultura y Artes).
Una vez determinados los sectores asignamos las observaciones seleccionadas y procesadas a su gran sector correspondiente, incorporando el valor del coeficiente previamente procesado por el modelo de regresión. Hecho esto se suman todos los coeficientes de cada grupo y se eliminan los nombres particulares para visualizar los cinco grandes sectores.

Análisis de los Resultados obtenidos

Resultados de la Regresión Logística

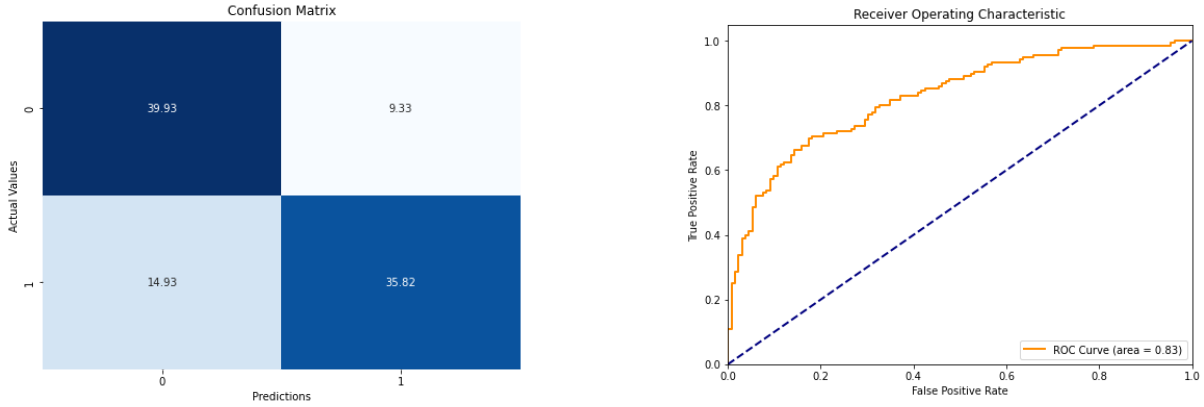

Figura 5

The Lasso model has an AUC of 0.83, AIC of -867.49, and BIC of 583.27

El análisis de los resultados del modelo de regresión logística con penalización Lasso muestra varios elementos clave sobre su rendimiento y efectividad. A continuación, se profundiza en cada uno de los aspectos del modelo basándose en los gráficos de la matriz de confusión y la curva ROC, junto con las métricas de evaluación reportadas.
La matriz recoge que que un 39.93% de las predicciones del modelo arrojan un No Default predicho que se corresponde con un No Default real (Verdaderos Negativos), un 9.33% arrojan un Default predicho que no se corresponde con un Default real (Falsos Positivos), un 14.93% arrojan un No Default predicho que no se corresponde con un No Default real (Falsos Negativos) y un 35.82% arrojan un Default predicho que se corresponde con un Default real (Verdaderos Positivos).

La precisión indica qué tan bien el modelo clasifica a las empresas en las categorías de default y no default en general y una alta precisión es deseable, ya que implica que el modelo es efectivo en la clasificación correcta de las empresas.

$$Precisión\ (accuracy)\ = \frac{Verdaderos\ Positivos\ (VP) + Verdaderos\ Negativos(VN)}{Total\ de\ Predicciones}$$

El modelo clasifica correctamente el 75.75% de las veces, ya sea como default o no default.

La sensibilidad mide la capacidad del modelo para identificar correctamente a las empresas que harán default. Esta métrica es crucial en escenarios financieros donde el coste de no detectar un default puede ser significativamente alto, como pérdidas económicas graves o riesgos de crédito. Una alta sensibilidad significa que el modelo es efectivo en capturar la

mayoría de los defaults, reduciendo el riesgo de pérdidas inesperadas.

$$Sensibilidad\ (recall\ o\ tasa\ de\ verdaderos\ positivos) = \frac{Verdaderos\ Positivos\ (VP)}{Verdaderos\ Positivos\ (VP) + Falsos\ Negativos\ (FN)}$$

El modelo identifica correctamente el 70.58% de las empresas que realmente hacen default.

La especificidad evalúa cuán bien el modelo puede identificar a las empresas que no harán default. Una alta especificidad implica que el modelo es confiable en reconocer a las empresas seguras, lo cual es importante para no restringir injustamente el acceso al crédito a empresas viables.

$$Especificidad\ (tasa\ de\ verdaderos\ negativos) = \frac{Verdaderos\ Negativos\ (VN)}{Verdaderos\ Negativos\ (VN) + Falsos\ Positivos\ (FP)}$$

El modelo identifica correctamente el 81.06% de las PYMEs que no hacen default.

El Valor Predictivo Positivo indica la probabilidad de que una empresa clasificada como en riesgo de default realmente termine en default. Un VPP alto es vital para asegurar que los recursos se asignen eficientemente, como la implementación de medidas de seguimiento o recuperación, y para no tomar acciones innecesarias o costosas contra empresas que son clasificadas incorrectamente como de alto riesgo.

$$Valor\ Predictivo\ Positivo\ (VPP) = \frac{Verdaderos\ Positivos\ (VP)}{Verdaderos\ Positivos\ (VP) + Falsos\ Positivos\ (FP)}$$

Si el modelo predice un default, hay un 79.33% de probabilidad de que la empresa realmente haga default.

El Valor Predictivo Negativo refleja la probabilidad de que una empresa clasificada como no en riesgo de default realmente no lo esté. Un VPN alto es esencial para minimizar los falsos negativos, es decir, asegurarse de que las empresas que se consideran sin riesgo realmente estén libres de default. Esto ayuda a evitar sorpresas desagradables y a gestionar mejor el riesgo de crédito.

$$Valor\ Predictivo\ Negativo\ (VPN) = \frac{Verdaderos\ Negativos\ (VN)}{Verdaderos\ Negativos\ (VN) + Falsos\ Negativos\ (FN)}$$

Si el modelo predice que no habrá default, hay un 72.78% de probabilidad de que la empresa realmente no haga default.

La Curva del Receptor Operativo (ROC) y el Área Bajo la Curva (AUC) son indicadores críticos de la capacidad del modelo para discriminar entre las clases positivas y negativas. La curva ROC traza la tasa de verdaderos positivos contra la tasa de falsos positivos a varios umbrales de decisión.

El AUC de 0.83 indica un rendimiento bastante bueno, aunque no excelente. Un AUC de 1 sería perfecto, mientras que un AUC de 0.5 indicaría un rendimiento no mejor que el azar. Un AUC de 0.83 sugiere que el modelo tiene una capacidad considerablemente buena para distinguir entre las empresas que entrarán en default y las que no.

Las métricas de información, específicamente el Criterio de Información Akaike (AIC) y el Criterio de Información Bayesiano (BIC), ayudan a evaluar la calidad del modelo considerando la complejidad del modelo y el ajuste a los datos.
El AIC es de -867.49. Que este valor sea bajo sugiere un mejor modelo. Aunque el AIC es negativo aquí, lo importante es que en comparaciones relativas, valores más bajos son mejores, indicando un equilibrio favorable entre complejidad del modelo y ajuste a los datos.
El BIC es de 583.27. Similar al anterior, un valor más bajo indica un mejor modelo. Sin embargo, el BIC penaliza modelos más complejos más fuertemente que el AIC.

El modelo muestra una capacidad buena para identificar correctamente los casos de default, como lo demuestra la AUC de 0.83. Sin embargo, hay una cantidad significativa de falsos negativos y falsos positivos que podrían ser críticos dependiendo del costo asociado a estas decisiones erróneas en un contexto real.
Desde una perspectiva más amplia, el modelo equilibra bien la complejidad y el rendimiento general, como lo indican los valores de AIC y BIC. Sin embargo, podría beneficiarse de una optimización adicional, posiblemente mediante una mejor selección de características o ajuste de parámetros.

Resultados de la Support Vector Machine

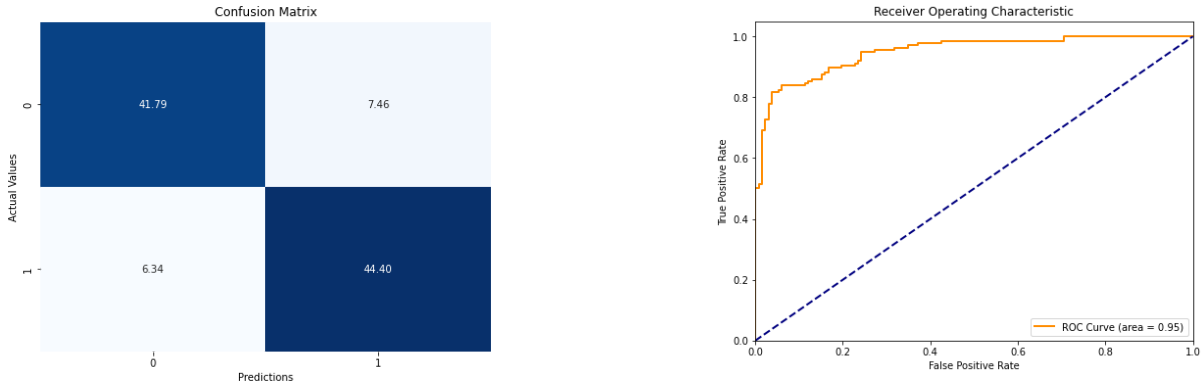

Figura 6

El mejor valor de C es: 11, con una puntuación de validación cruzada de: 0.86048440173752.
The SVM model has an AUC of 0.95, AIC of 124.24, and BIC of 2286.01

Un *C* más alto puede llevar a un modelo que se ajusta muy bien al conjunto de entrenamiento pero pierde capacidad de generalización, mientras que un *C* más bajo podría ser demasiado general. Para la selección de este parámetro utilizamos validación cruzada que devolvió un óptimo de 11. Gracias a esto encontramos un equilibrio entre sesgo y varianza, crucial en predicciones financieras donde ambos los falsos positivos y los falsos negativos tienen implicaciones económicas significativas.

La matriz arroja un 41.79% de Verdaderos Negativos, un 7.46% de Falsos Positivos, un 6.34% de Falsos Negativos y un 44.40% de Verdaderos Positivos. Con esto vemos que el modelo tiene un 86.19% de precisión, un 87.50% de sensibilidad, un 84.85% de especificidad, un 85.62% de valor predictivo positivo y un 86.83% de valor predictivo negativo. Esto refleja que sensibilidad y especificidad están bastante bien balanceados y sus valores son buenos. Esto hace que también los valores predictivos tanto positivos como negativos sean muy similares y estén en la misma ventana de valores buenos, lo que es muy importante en el contexto de análisis de riesgo de crédito en el cual nos encontramos.

En lo que respecta a la Curva ROC vemos que genera un AUC de 0.95, que es un valor muy bueno y que desvela que el modelo tiene un muy excelente rendimiento en todos los umbrales de clasificación, lo que indica su capacidad para discriminar entre las clases positivas (casos de default) y negativas (casos de no default).

El modelo tiene un AIC de 124.24 y un BIC de 2286.01. Estos valores son altos debido a que estamos utilizando un kernel radial (‘rbf’), no se puede acceder directamente a los coeficientes del modelo de la misma manera que lo haríamos con un kernel lineal, por lo que utilizamos el número de vectores de soporte como un proxy para la complejidad del modelo. Aunque no es exactamente igual a tener los coeficientes, los vectores de soporte juegan un papel crítico en la definición del modelo. El modelo ajusta muchos vectores de soporte al entrenar, lo que hace que aumente la *k* de la fórmula del AIC y BIC y por tanto estos valores. Es por ello que tomaremos estos valores para comparar los modelos SVM con kernel radial entre sí y no utilizaremos estos dos criterios para compararlos con el resto de modelos.

Resultados del Z-Score de Altman

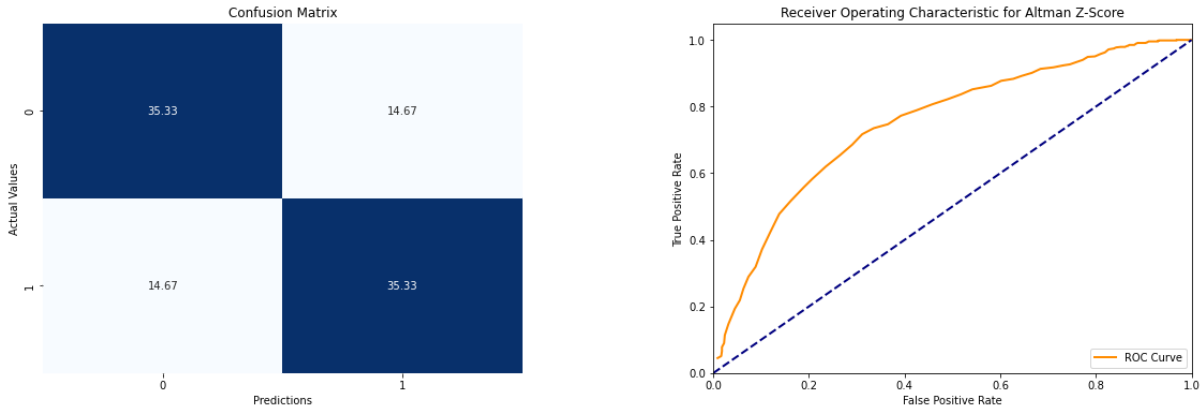

Figura 7

El AUC calculado es: 0.7141200383663809

La matriz de confusión del modelo Z-Score de Altman refleja un 35.33% de Verdaderos Negativos, un 14.67% de Falsos Positivos, un 14.67% de Falsos Negativos y un 35.33% de Verdaderos Positivos. Altman contemplaba tres zonas en las que clasificar la situación de la empresa respecto a su probabilidad de incumplimiento. Para la clasificación del modelo utilicé el umbral que separa la zona 'gris' de la zona 'default', por lo cual tanto 'grises' como 'seguros' quedan clasificados como No Default (0).
Con estos resultados obtenemos un 70.66% de precisión, un 70.66% de sensibilidad, un 70.66% de especificidad, un 70.66% del valor predictivo de los positivos y un 70.66% de valor predictivo de los negativos.
Dado que el Z-Score de Altman tradicionalmente utiliza un umbral fijo, la curva ROC y el AUC no resultan informativas como lo serían para un modelo con umbrales variables. Sin embargo, en este modelo de scoring vamos a dibujar la curva y obtener el AUC para diferentes potenciales umbrales, aunque vamos a respetar el determinado por el autor del modelo. La curva ROC señala un AUC de 0.71, que es aceptable pero no brillante a la hora de clasificar bajo diferentes umbrales.
El modelo tiene un rendimiento entre aceptable y bueno. Es decentemente efectivo para clasificar correctamente las clases y de igual modo confiable en reconocer a las empresas seguras.

Resultados del Score de Elisabetsky

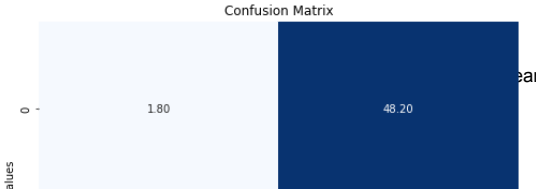

Figura 8

La matriz de confusión del modelo Score de Elisabetsky refleja un 1.80% de Verdaderos Negativos, un 48.20% de Falsos Positivos, un 1.20% de Falsos Negativos y un 48.80% de Verdaderos Positivos. Con ello obtenemos un 50.60% de precisión, un 97.60% de sensibilidad, un 3.6% de especificidad, un 50.31% del valor predictivo de los positivos y un 60.00% de valor predictivo de los negativos.

Como se puede ver el modelo es deficiente para la clasificación en este contexto ya que vemos que clasifica a prácticamente todas las empresas como potenciales Default, lo que lo hace inutilizable. Esto se puede dar por los siguientes factores:

- Los datos que componen el dataset quizás no son los que la teoría requiere específicamente para construir el modelo. Aunque en la formulación teórica de los ratios que se utilizan encontramos equivalentes en la base de datos que utilizamos, un desbalance tan pronunciado hacia la clasificación en una clase invita a pensar que alguno de los datos no se corresponde con el que realmente el modelo teórico solicita.
- Cabe la duda de cómo se comportaría el modelo con otro umbral. Si realmente los datos que trabajamos son los adecuados, el problema podría ser el umbral y se debería investigar cómo funciona un umbral mayor que sea más exigente con los requisitos para considerar a una empresa solvente.
- Kassai y Kassai (1998) destacan que los modelos de previsión de insolvencia fueron desarrollados a partir de una determinada muestra recogida en sus respectivas épocas y por eso mismo pueden no tener la misma eficacia actualmente si comparadas a la época de su desarrollo. Desde el desarrollo del modelo en 1976 han ocurrido recesiones, cambios en casas de interés, etc. que pueden afectar el comportamiento del crédito y no ser reflejados actualmente en este modelo. Además de remarcar que la base de datos comprende un horizonte temporal especialmente coyuntural como lo es el de la crisis de 2008.
- El comportamiento de los prestatarios puede cambiar con el tiempo debido a nuevas tecnologías, cambios en el mercado laboral, entre otros factores.

Resultados del Termómetro de Kanitz

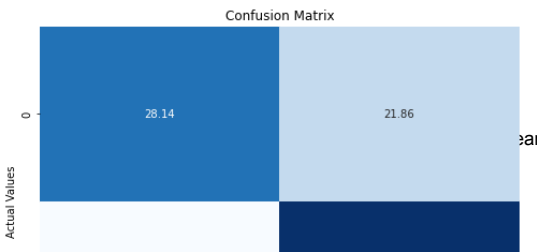

Figura 9

La matriz de confusión del modelo del Termómetro de Kanitz refleja un 28.14% de Verdaderos Negativos, un 21.86% de Falsos Positivos, un 18.56% de Falsos Negativos y un 31.44% de Verdaderos Positivos. Kanitz contemplaba un termómetro sobre el cual clasificar las empresas con respecto a la probabilidad de Default. En este caso, al igual que con el Z-Score de Altman, decidimos emplear el umbral que separa a las empresas en la zona 'gris' de las empresas en la zona 'default'. Por ello quedan clasificadas como No Default (0) las empresas de la zona 'segura' y las empresas de la zona 'gris'.
Con estos resultados obtenemos un 59.88% de precisión, un 62.88% de sensibilidad, un 56.28% de especificidad, un 58.98% del valor predictivo de los positivos y un 60.26% de valor predictivo de los negativos.
Este modelo tiene una capacidad discriminativa bastante pobre y prácticamente inexistente. Su precisión es poco mejor que una selección aleatoria lo que hace que la cantidad de errores sea demasiado significativa.

Resultados del impacto de los sectores en la probabilidad de Default

Sector General	Suma de Coeficientes pertenecientes al sector

Sector Primario	-0.414936
Sector Secundario	13.222377
Sector Terciario	12.869550
Sector Cuaternario	-1.589377
Sector Quinario	0.988125

Tabla 2

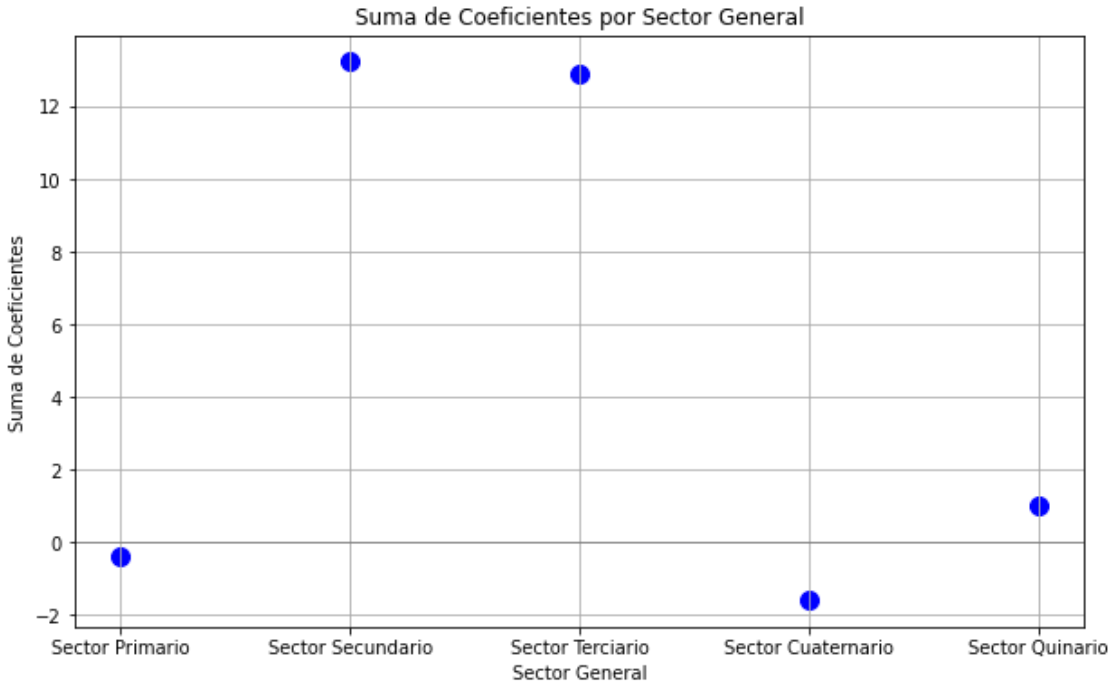

Figura 10

Las PYMES del sector primario (Agricultura, Ganadería, Pesca, Silvicultura, Minería) no tienen un impacto significativo en la probabilidad de default según el modelo de regresión logística. Esto sugiere que estar en el sector primario no aumenta ni disminuye la probabilidad de default en comparación con otros sectores.
La demanda de productos agrícolas y alimentarios tiende a ser menos elástica durante las recesiones económicas. El valor añadido bruto del sector primario mostró una mayor estabilidad en comparación con otros sectores durante la crisis.

Las PYMES del sector secundario (Industria Manufacturera, Construcción, Industria Química, Industria Metalúrgica, Industria Automotriz) tienen un coeficiente positivo alto. Esto indica que estar en el sector secundario aumenta significativamente la probabilidad de default. La razón puede ser la alta exposición a la crisis económica de 2008-2011, que afectó fuertemente a la industria manufacturera y la construcción.

La burbuja inmobiliaria que estalló en 2008 llevó a una caída significativa en la construcción, y la demanda de productos manufacturados también disminuyó. El Instituto Nacional de Estadística (INE) reportó que el índice de producción industrial cayó un 25% entre 2008 y 2011 . Además, el sector de la construcción vio una reducción del empleo del 50% durante el mismo periodo.

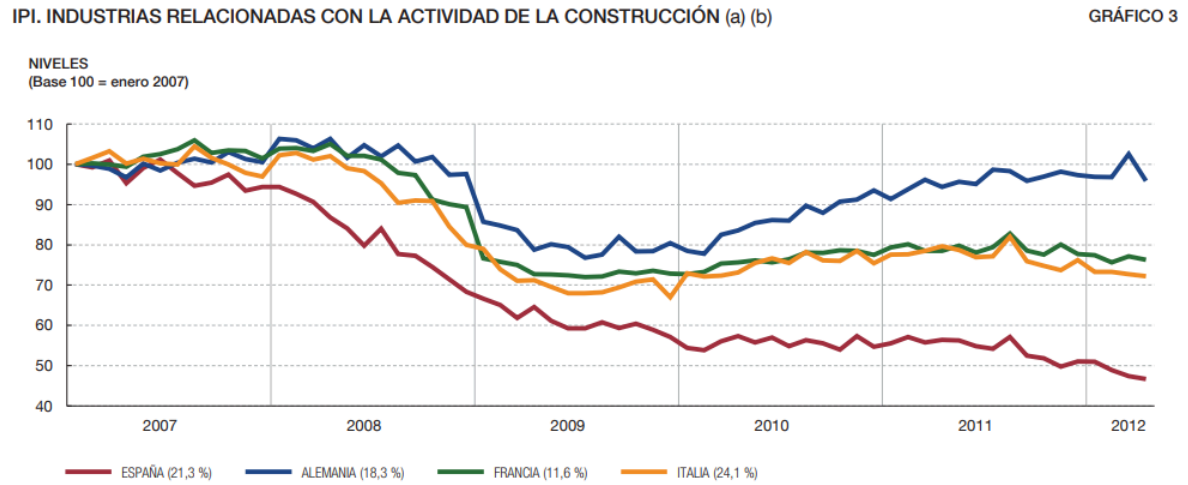

Figura 11

Fuente: Banco de España, 2014

Similar al sector secundario, las PYMES del sector terciario (Comercio, Transporte y Logística, Hostelería y Turismo, Educación, Salud, Finanzas y Seguros, Telecomunicaciones, Tecnologías de la Información) también tienen un coeficiente positivo alto. Esto sugiere que estar en el sector terciario también aumenta significativamente la probabilidad de default, posiblemente debido a la reducción del consumo y la actividad económica durante la crisis.

El turismo, una parte importante de la economía española, experimentó una disminución en el número de visitantes y en el gasto turístico. Según el INE, el turismo en España cayó significativamente durante 2008-2009, con una reducción del gasto turístico del 5.6%.

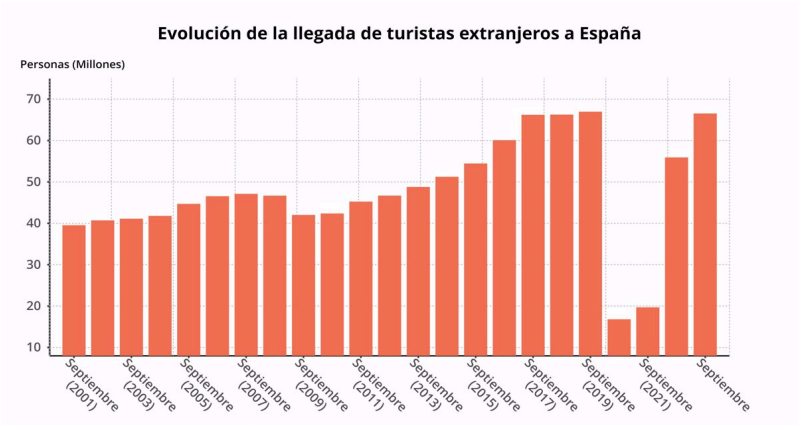

Figura 12

Fuente: Europapress, 2023

Las PYMES del sector cuaternario (Investigación y Desarrollo, Tecnología de la Información y Comunicación, Consultoría y Asesoría, Educación) tienen un coeficiente negativo. Esto

indica que estar en el sector cuaternario reduce la probabilidad de default. Este sector pudo haber sido menos afectado por la crisis o incluso haber visto oportunidades de crecimiento en tecnología y consultoría.
La demanda de servicios de TIC y consultoría se mantuvo e incluso creció en algunos casos, ya que las empresas buscaron optimizar operaciones y reducir costos. El sector TIC en Europa, incluido España, continuó creciendo. Además, los servicios de consultoría vieron una demanda sostenida debido a la necesidad de reestructuración empresarial .

Las PYMES del sector quinario (Gobierno y Administración Pública, ONGs y Organizaciones Internacionales, Servicios de Alta Dirección, Cultura y Artes) tienen un coeficiente positivo pero bajo. Esto sugiere un leve aumento en la probabilidad de default, posiblemente debido a la estabilidad relativa de este sector en comparación con otros durante la crisis.
El gasto público en España se mantuvo relativamente alto debido a las medidas de estímulo económico, aunque con ajustes significativos en ciertos sectores. Ejemplo de ello fue el 'Plan E', el cual incluyó dos fondos para apresuradas obras en municipios. En el primero, en el Fondo Estatal de Inversión Local (FEIL) realizado en 2009, se invirtieron 7.860 millones; en el segundo, llamado Fondo Estatal para el Empleo y la Sostenibilidad Local, en 2010, otros 4.250 millones de euros. En total, 12.110 millones de euros. (El Mundo, 2013)

- **Sectores Más Afectados (Secundario y Terciario):** Las PYMES en estos sectores tuvieron un alto riesgo de default, probablemente debido a la disminución en la demanda y el impacto directo de la crisis económica.
- **Sectores Menos Afectados (Cuaternario):** Las PYMES en el sector cuaternario, que incluye tecnología y consultoría, pudieron haber encontrado oportunidades de crecimiento o haber sido menos afectadas por la crisis.
- **Impacto Neutral (Primario):** Las PYMES en el sector primario no muestran un impacto significativo en la probabilidad de default.

Conclusiones

En esta última sección extraeremos las conclusiones derivadas del análisis de los resultados de los modelos de scoring y machine learning y de la influencia de los sectores de actividad en la probabilidad de default basado en los valores de los coeficientes calculados mediante el modelo de regresión logística.

Modelo	Precisión	Sensibilidad	Especificidad	VPP	VPN	AUC	AIC	BIC
RL	75,75%	70,58%	81,06%	79,33%	72,78%	0,83	-867,49	583,27
SVM	86,19%	87,50%	84,85%	85,62%	86,83%	0,95	124,24	2286,01
Altman	70,66%	70,66%	70,66%	70,66%	70,66%			
Elisabetsky	50,60%	97,60%	3,60%	50,31%	60,00%			
Kanitz	59,88%	62,88%	56,28%	58,98%	60,26%			

Tabla 3

A raíz de los resultados obtenidos los modelos se puede concluir que los modelos machine learning seleccionados son más precisos y completos que los modelos de scoring seleccionados, ya que la diferencia entre las métricas de precisión, sensibilidad, especificidad, valor de las predicciones positivas y valor de las predicciones negativas es notable. Cabe discutir si la mejora que supone la regresión logística con respecto del Z-Score de Altman justifica la pérdida de interpretabilidad debido al mayor número de variables de la primera. Sin embargo, tanto el Score de Elisabetsky y el Termómetro de Kanitz presentan métricas demasiado poco consistentes como para poder plantearse como alternativa ante los modelos de machine learning. Esto se puede deber a que son modelos antiguos o a que la base de datos usada en este caso es representativa de un periodo de tiempo caracterizado por un entorno macroeconómico mundial atípico, lo que altera los balances normales de las empresas.

Pese a tener un mejor rendimiento, los modelos de machine learning pierden interpretabilidad en comparación, ya que aunque los modelos de scoring utilizan en torno a 7 u 8 variables cada uno, los modelos machine learning utilizan 28 variables seleccionadas.Gracias a una evaluación de coeficientes en el modelo de regresión o una evaluación de importancia en el SVM, se pueden identificar aquellas variables más relevantes de un espectro de variables que, si bien es más amplio que el de los modelos de scoring, tampoco resulta en demasiado sobreajuste, sobre todo gracias a la aplicación de la penalización Lasso en el modelo de regresión. El modelo menos interpretable es la SVM, debido a su mayor Criterio de Akaike y Criterio Bayesiano, que se debe principalmente a la complejidad que aumenta con el número de vectores de soporte (support vectors) y con la elección de un kernel complejo como es RBF. Es importante recalcar que las SVMs en sí mismas no son modelos muy interpretables en términos de comprensión directa de cómo se toman las decisiones de clasificación

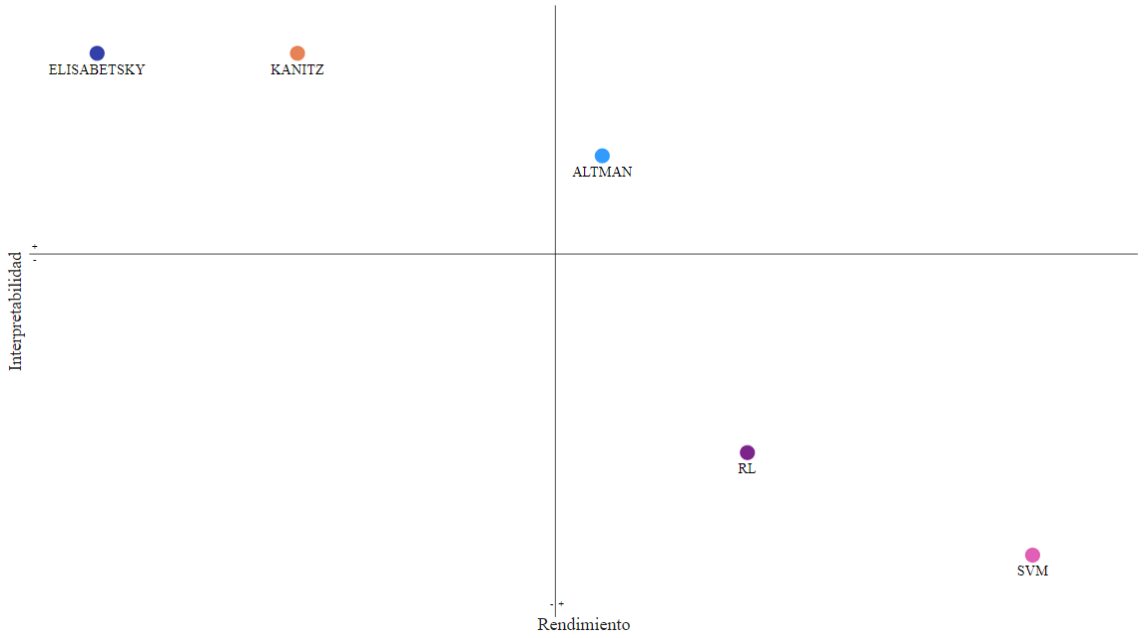

Figura 13

En este gráfico se mide el rendimiento de los modelos en comparación con su interpretabilidad. La escala del eje de rendimiento representa la media de los valores de la precisión, sensibilidad, especificidad, valor de las predicciones positivas y valor de las predicciones negativas y para la escala del eje de interpretabilidad se tienen en cuenta los criterios AIC y BIC para los modelos machine learning y la cantidad de variables utilizadas para los modelos de scoring.
El equilibrio óptimo entre rendimiento e interpretabilidad es un factor de decisión determinado por la finalidad del uso del modelo y por los criterios de la entidad que lo utilice. Si hubiera que destacar un modelo que funciona mejor que la media de los evaluados en ambos aspectos sería el Z-Score de Altman, que parece el más equilibrado y adecuado para satisfacer ambas necesidades.

Los resultados del análisis sectorial indican que la probabilidad de default en las PYMES varía significativamente según el sector económico en el que operan, reflejando cómo diferentes factores externos, como la crisis económica, pueden impactar a las industrias de manera desigual. El sector secundario y terciario mostraron una marcada vulnerabilidad durante la crisis de 2008, evidenciado por altos coeficientes positivos en el modelo de regresión logística, lo que sugiere una significativa probabilidad de default comparado con otros sectores. Esto se debe, en gran medida, a su alta exposición a las fluctuaciones del mercado y la demanda, particularmente afectadas por la recesión económica y la contracción en el consumo y la inversión.
Por otro lado, el sector cuaternario presentó un coeficiente negativo, indicando una menor probabilidad de default. Este sector parece haberse beneficiado de la necesidad de servicios tecnológicos y de consultoría durante la crisis, donde empresas buscaron adaptarse y optimizar recursos frente a las nuevas realidades económicas. Mientras tanto, el sector primario y quinario mostraron impactos menos significativos en la probabilidad de default,

sugiriendo una estabilidad relativa durante la crisis. Esto puede atribuirse a la naturaleza esencial de los bienes y servicios ofrecidos por estas áreas, como los productos agrícolas y los servicios gubernamentales, que tienden a ser menos sensibles a las condiciones económicas adversas.

En conclusión, este trabajo ha demostrado que la adopción de modelos avanzados de machine learning ofrece un aumento significativo en la precisión de la predicción de defaults en comparación con los modelos de scoring tradicionales, aunque a costa de una reducción en la interpretabilidad. Este desequilibrio destaca la importancia de una selección cuidadosa del modelo basada en el contexto específico de uso y los valores de la entidad que los implementa. Además, el análisis del impacto de los sectores económicos en la probabilidad de default ha revelado diferencias significativas en cómo distintos sectores responden a las crisis económicas, proporcionando una comprensión más matizada de los riesgos sectoriales.

Referencias Bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723

Alonso, A., & Carbó, J. M. (2020). Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost. Documentos de Trabajo (N.º 2032). Banco de España, Eurosistema

Alonso, A., & Carbó, J. M. Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation. Summary of Banco de España Working Paper no. 2105

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(4), 589-609. https://www.sciencedirect.com/science/article/abs/pii/S1366554522000096

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609

Banco Central Europeo. (2024). ECB guide to internal models. Banco Central Europeo. Recuperado de https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guides202402_internalmodels.bg.pdf

Banco de España. (2014). *Informe sobre la crisis financiera y bancaria en España, 2008-2014*. Banco de España.

Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., & Tarantino, B. (2023). Machine learning and credit risk Empirical evidence from small- and mid-sized businesses. Socio-Economic Planning Sciences, 90, 101746

Columbus, L. (2021, January 17). 76% of enterprises prioritize AI & machine learning in 2021 IT budgets. Forbes. Updated January 26, 2021. Recuperado de https://www.forbes.com/sites/louiscolumbus/2021/01/17/76-of-enterprises-prioritize-ai--machine-learning-in-2021-it-budgets/#:~:text=43%25%20of%20enterprises%20say%20their,be%20significantly%20increasing%20their%20budgets

Comisión Europea. (2003). Recomendación de la Comisión, de 6 de mayo de 2003, sobre la definición de microempresas, pequeñas y medianas empresas [notificada con el número C(2003) 1422]. Diario Oficial de la Unión Europea, núm. 124, pp. 36-41

Dimitrescu, E.-I., Hurlin, C., Hué, S., & Tokpavi, S. (2020, Enero 1). Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. SSRN Electronics Journal. https://www.researchgate.net/publication/340504937

Elisabetsky, R. (1976). Um modelo matemático para decisões de crédito no banco comercial. Dissertación de maestría – Escuela Politécnica, Universidad de São Paulo, São Paulo

El Mundo. (2009, febrero 12). España entra en recesión por primera vez en 15 años. Recuperado de https://www.elmundo.es/mundodinero/2009/02/12/economia/1234425902.html

El Mundo. (2014, abril 21). La morosidad del sistema financiero vuelve a bajar tras el cambio del sistema de cálculo. El Mundo. Recuperado de https://www.elmundo.es/economia/2014/04/21/5354d75d268e3eae218b4570.html

Europapress. (2023, septiembre 29). España recibe casi un 20% más de turistas hasta septiembre anticipando un nuevo récord a fin de año. epturismo. Recuperado de https://www.europapress.es/turismo/nacional/noticia-espana-enfila-recta-final-ano-record-numero-turistas-gasto-realizan-20231102092506.html

European Banking Authority. (2021, Noviembre 11). EBA Discussion Paper on Machine Learning for IRB Models

European Banking Authority. (2023, Agosto). Machine Learning for IRB Models [Follow-Up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models]

Fernández, A. (2019). Inteligencia artificial en los servicios financieros. Boletín Económico 2/2019, Artículos Analíticos, 29 de marzo de 2019. Banco de España, Eurosistema

Fernández, R. (2024). Evolución de la ratio de préstamos dudosos (NPL) en España desde 2005 a 2022. Statista. Recuperado de https://www.statista.com/estadisticas/1132704/evolucion-ratio-prestamos-dudosos-espana/

Instituto de Contabilidad y Auditoría de Cuentas (ICAC). (2013). NIA-ES 705: Opinión modificada en el informe emitido por un auditor independiente (adaptada para su aplicación en España mediante Resolución del 15 de octubre de 2013)

Instituto Nacional de Estadística (INE). (2009). Clasificación Nacional de Actividades Económicas (CNAE-2009)

Kassai, J. R., & Kassai, S. (1998). Desvendando o termômetro de Kanitz. En Anales del 22º Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração (ANPAD), Foz do Iguaçu, Brasil

Kanitz, S. C. (1976). Indicadores contábeis financeiros – previsão de insolvência: a experiência da pequena e média empresa brasileira. Tesis de libre-docencia entregada al Departamento de Contabilidad de la FEA/USP

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109-131

Pérez, R. (2013, mayo 19). ¿Se acuerdan del «Plan E»?: 12.000 millones de euros hechos trizas. ABC España. Recuperado de https://www.abc.es/espana/20130519/abci-acuerdan-plan-millones-euros-201305181953.html

Schwarz, G. E. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2002). Credit scoring and its applications. Society for Industrial and Applied Mathematics

Anexo

Librerías de Python utilizadas para la investigación:

- Librería itertools
- De imblearn.combine SMOTEENN
- De imblearn.over_sampling RandomOverSampler
- Librería matplotlib
- Librería numpy
- Librería pandas
- De scipy.stats pointbiserialr
- Librería seaborn
- Librería shap
- De sklearn.compose ColumnTransformer
- De sklearn.datasets make_classification
- De sklearn.decomposition PCA
- De sklearn.ensemble RandomForestClassifier
- De sklearn.feature_selection RFECV
- De sklearn.inspection permutation_importance
- De sklearn.linear_model LogisticRegression, LogisticRegressionCV
- De sklearn.metrics classification_report, confusion_matrix, roc_curve, auc, log_loss, accuracy_score, recall_score, precision_recall_curve, f1_score
- De sklearn.model_selection train_test_split, cross_val_score, GridSearchCV
- De sklearn.pipeline Pipeline
- De sklearn.preprocessing StandardScaler, OneHotEncoder
- De sklearn.svm SVC
- De sklearn.utils resample
- Librería statsmodels

Anexo de tablas y figuras

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, [Nombre completo del estudiante], estudiante de [nombre del título] de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "[Título del trabajo]", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir "no he usado ninguna"]:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
5. **Interpretador de código:** Para realizar análisis de datos preliminares.
6. **Estudios multidisciplinares:** Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
8. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
9. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
10. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.

11. **Generador de datos sintéticos de prueba:** Para la creación de conjuntos de datos ficticios.

12. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.

13. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

14. **Generador de encuestas:** Para diseñar cuestionarios preliminares.

15. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: [Fecha] MIELGO LARRIBA LUIS - 39466701N

Firmado digitalmente por MIELGO LARRIBA LUIS - 39466701N
Fecha: 2024.06.21 19:09:11 +02'00'

Firma: _____