Universidad Anáhuac Norte

Maestría en Tecnologías de Información e Inteligencia Analítica

Redes Neuronales y Máquinas de Soporte

Modelo de Riesgo Crediticio

Profesor: Dr. Román Rodríguez Aguilar

Alumno: Lic. André Marx Puente Arévalo

Marzo 2023

Índice

Resumen

El presente trabajo aspira aplicar, al menos, un par de metodologías de los temas vistos en la materia de “Redes Neuronales y Máquinas de Soporte Vectorial”, con el objetivo de predecir la probabilidad de impago, para lo cual, se construyó un conjunto de datos sobre la información resumida de las facturas emitidas y recibidas de los clientes de una Fintech mexicana dedicada a brindar servicios financieros. Se implementaron dos metodologías: Máquinas de Soporte Vectorial y Redes Neuronales. Tras analizar y procesar el conjunto de datos, se usaron métodos de muestreo para el balanceo de clases y búsqueda de hiperparámetros, se ajustó una red neuronal que lograra clasificar entre pagadores y morosos.

Palabras Clave: *Clasificación, morosidad, redes neuronales, máquinas de soporte vectorial*

Introducción

Desde hace mucho tiempo, las instituciones financieras han sido el principal canal a través del cual las personas y las empresas pueden acceder a los servicios financieros necesarios. Estas instituciones han sido fundamentales en la intermediación entre los prestamistas y los prestatarios, el ahorro y la inversión, y en general en la gestión de las finanzas a nivel global.

Sin embargo, con el rápido avance de la tecnología y el surgimiento de nuevas empresas y modelos de negocio, las instituciones financieras han tenido que evolucionar para mantenerse al día con las demandas y necesidades de los consumidores. Es aquí donde entran en juego las fintech.

Las fintech son empresas que utilizan la tecnología para ofrecer servicios financieros más eficientes, convenientes y accesibles que las instituciones financieras tradicionales. Estas empresas han logrado atraer a una gran cantidad de clientes en todo el mundo gracias a sus modelos de negocio innovadores y sus productos y servicios personalizados. Uno de los aspectos clave de las fintech es su capacidad para aprovechar la gran cantidad de datos que se generan en el ámbito financiero para tomar decisiones más informadas.

En primer lugar, pueden utilizar los datos para entender mejor a sus clientes. Al recolectar información sobre sus patrones de gasto, sus preferencias de inversión y su comportamiento financiero en general, pueden crear perfiles más precisos de sus clientes y ofrecer servicios y productos personalizados para satisfacer sus necesidades específicas.

Además, las fintech pueden utilizar los datos para mejorar la gestión de riesgos. Al analizar grandes cantidades de información sobre el historial crediticio, la solvencia y el

comportamiento de pago de los clientes, pueden tomar decisiones más precisas y reducir el riesgo de impago o de fraude.

Por lo que, el presente proyecto se enfocará en generar un modelo de riesgo crediticio para una Fintech mexicana, de reciente creación, que se enfoca a brindar servicios financieros a Pequeñas y Medianas Empresas (PyMEs) con el objetivo de impulsar su crecimiento. El estado actual de la Fintech es que ya cuenta con un modelo de "scoring" pero basado en reglas de negocio, debido a que, no se contaba con los suficientes datos para realizar un modelo más sofisticado.

Se llevará a cabo la evaluación y comparación de diferentes modelos de clasificación con el objetivo de determinar cuál es el mejor para predecir el comportamiento de impago utilizando información de las facturas de clientes. Para ello, se trabajará con datos resumidos de las facturas emitidas y recibidas a lo largo del tiempo.

Material y Métodos

Prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis es una prueba no paramétrica que se utiliza para determinar si hay una diferencia estadísticamente significativa entre tres o más grupos independientes en una variable ordinal o continua. Esta prueba se basa en los rangos de las observaciones de cada grupo y se utiliza cuando los datos no cumplen con los supuestos de normalidad y homogeneidad de varianza que se requieren en otras pruebas paramétricas (Conover, 1999). Las hipótesis a probar son:

H_0: Las *k* funciones de distribución son idénticas

vs

H_a: Al menos una de las *k* poblaciones se distribuye diferente

Para realizar la prueba de Kruskal-Wallis, se calcula una estadística de prueba llamada H que mide la diferencia en los rangos medios de los grupos. Si H es significativo, se concluye que hay diferencias significativas entre los grupos. Otra manera de concluir es utilizando el *p-value* y la zona de rechazo, donde si $p - value < \alpha$, α es el nivel de significancia al 0.05, se rechaza la hipótesis nula, en caso contrario se acepta.

Muestreo

El muestreo es una técnica utilizada en estadística para seleccionar una muestra representativa de una población. Su objetivo es estimar las características de la población a partir de los datos obtenidos en la muestra.

Una problemática muy común al trabajar con datos es el desequilibrio de clases, el cual, ocurre cuando una clase o categoría de interés tiene una proporción significativamente menor de ejemplos que otra clase o categoría en un conjunto de datos. Los métodos más utilizados para dar solución a esto son los que se muestran en la Figura 1.

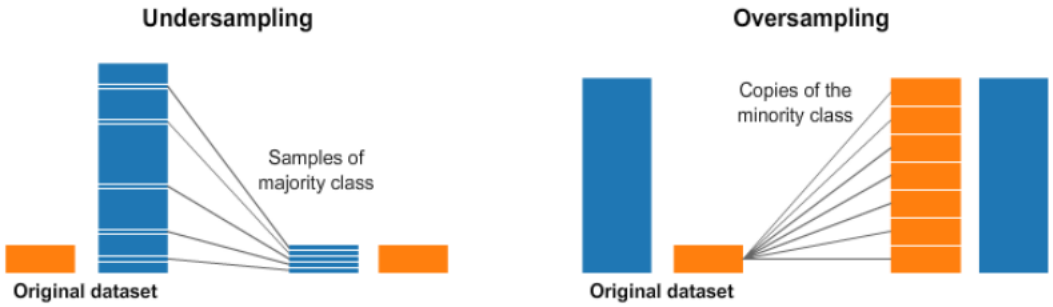

Figura1. Ejemplo de cómo funciona el método “undersample” y “oversample” (Badr, 2019)

- *Undersample*: Su objetivo es reducir el tamaño de la muestra de la clase mayoritaria de manera que su tamaño sea comparable al de la clase minoritaria, lo que puede mejorar la capacidad del modelo para reconocer y clasificar correctamente los ejemplos de la clase minoritaria.
- *Oversample*: Su objetivo es aumentar el tamaño de la muestra de la clase minoritaria para que sea comparable al de la clase mayoritaria, lo que puede mejorar la capacidad del modelo para reconocer y clasificar correctamente los ejemplos de la clase minoritaria.

Máquinas de Soporte Vectorial

El algoritmo de Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) es un método de aprendizaje supervisado utilizado para la clasificación y regresión. En este caso, por el tipo de modelo que se busca crear, me enfocaré en explicar su funcionamiento para la clasificación.

El objetivo de SVM es encontrar un hiperplano que maximice la distancia entre los puntos de las clases distintas en el espacio de características, es decir, aquel que separa de manera óptima las dos clases. Como se puede apreciar en la Figura 2. Un hiperplano es un subespacio de dimensión d-1 en un espacio vectorial de dimensión d. En el caso de datos que no son linealmente separables, SVM utiliza el método del kernel para mapear los datos a un espacio de características de mayor dimensión en el cual sí sean linealmente separables (Hastie, Tibshirani, & Friedman, 2009).

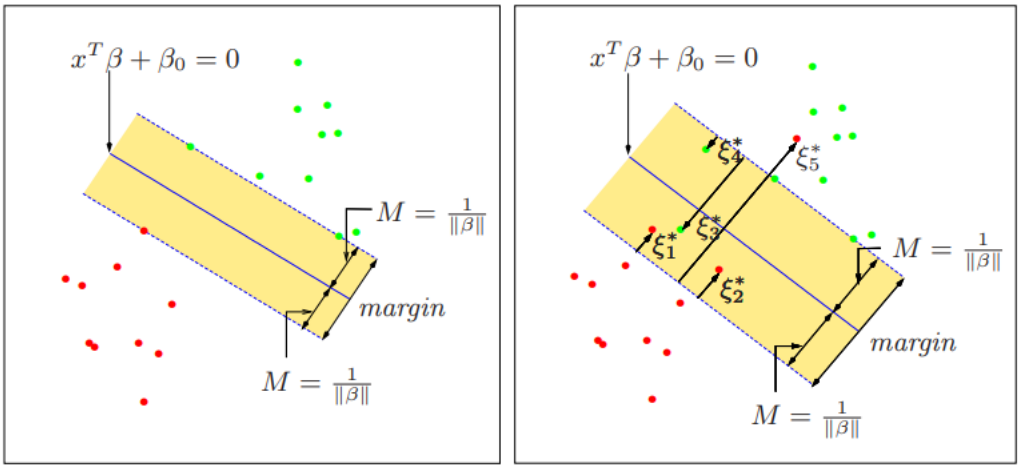

Figura 2. Ejemplo de Hiperplano de separación y el margen (Hastie, Tibshirani, & Friedman, 2009)

El margen es la distancia perpendicular entre el hiperplano separador y el hiperplano que pasa sobre los puntos más cercanos (podría no existir), los vectores soporte como se puede ver en la Figura 2.

Redes Neuronales

Una red neuronal es un modelo computacional inspirado en el cerebro humano que se utiliza en el campo del aprendizaje profundo (deep learning) para resolver tareas de clasificación, regresión, reconocimiento de imágenes, procesamiento del lenguaje natural, entre otros.

El funcionamiento de una red neuronal se basa en la interconexión de nodos (neuronas) que reciben una entrada y generan una salida. Cada neurona está asociada a un peso que se ajusta durante el proceso de entrenamiento para mejorar el rendimiento del modelo.

En el caso de la clasificación, la red neuronal se compone de una capa de entrada, una o varias capas ocultas y una capa de salida. Como se aprecia en la Figura 3. La capa de entrada recibe la entrada del modelo, que puede ser una imagen, un texto, un audio, etc. La capa oculta procesa la información de entrada mediante la aplicación de operaciones lineales y no lineales para extraer características relevantes. La capa de salida genera la salida del modelo, que puede ser una probabilidad o una etiqueta de clase (Goodfellow, Bengio, & Courville, 2016).

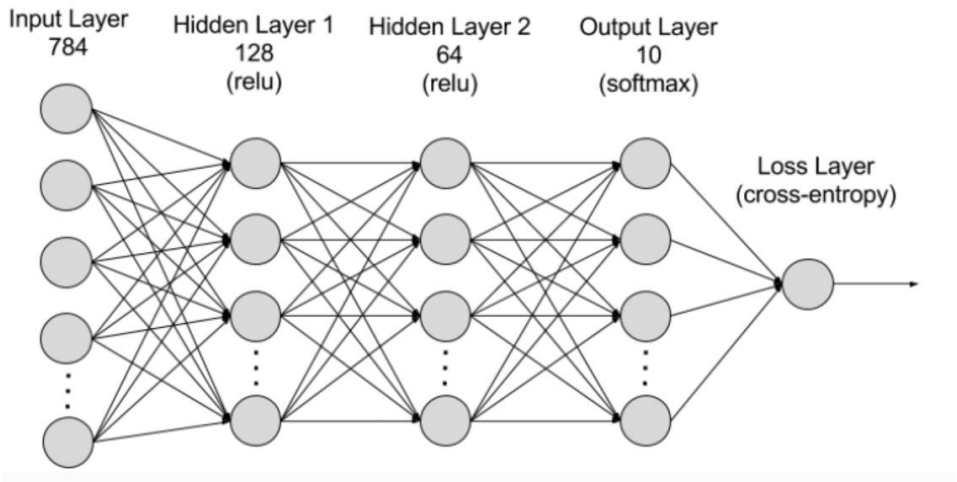

Figura 3. Arquitectura de una red neuronal profunda (AWS, 2023)

Existen varios tipos de redes neuronales, cada una diseñada para resolver un tipo específico de problema. A continuación, se describen algunos de los tipos más comunes:

- *Redes neuronales convolucionales (CNN)*: son utilizadas para tareas de procesamiento de imágenes y video. Se caracterizan por tener capas convolucionales y capas de pooling que permiten la detección de características locales en la imagen.
- *Redes neuronales recurrentes (RNN):* son utilizadas para tareas de procesamiento del lenguaje natural y series de tiempo. Se caracterizan por tener conexiones recurrentes que permiten el procesamiento de secuencias de datos.
- *Redes neuronales totalmente conectadas (FNN):* son utilizadas para tareas de clasificación y regresión en general. Se caracterizan por tener una o varias capas ocultas totalmente conectadas que permiten el procesamiento de características de alta dimensión.

El proceso de entrenamiento de una red neuronal se lleva a cabo mediante la retropropagación del error, que consiste en calcular la diferencia entre la salida del modelo y la salida esperada, y ajustar los pesos de la red para minimizar esta diferencia. Este proceso puede requerir un gran conjunto de datos etiquetados y un alto poder computacional para ajustar los millones de pesos de la red.

Resultados

El conjunto de datos seleccionado almacena información de las facturas de los clientes de la Fintech en cuestión. Se utilizaron dos fuentes de datos para construirla:

- Resumiendo información de las facturas utilizando reglas de negocio.
- Indicadores de riesgo que proporciona el proveedor de datos de facturación.

Por otro lado, para establecer si un cliente se encuentra en mora (más de 30 días de atraso), se utilizó información solamente de su primer crédito que le otorgaron.

Análisis Exploratorio de Datos

Primero se validó que el conjunto de datos no tuviera información repetida, por lo que se confirmó que se cuenta con información de 130 clientes únicos y cada uno cuenta con 23 variables mismas que se enlistan en el apéndice en la sección de "variables".

Se comprobó la completitud de las variables, es decir, cuál es su porcentaje de valores no nulos y se identificó lo siguiente:

Cuadro 1. Variables que tienen valores nulos

Variable	Completitud
blacklistst atus	0%
intercompanytransactions_value	76%
customerconcentration_value	99%
accountinginsolvency_value	68%
canceledissuedinvoices_value	99%

(Puente, 2023)

Dado que hay tres variables (blacklistedstatus, intercompanytransactions_value y accountinginsolvency_value) que exceden el tener 20% de valores faltantes, serán eliminadas, porque imputar valores a estas, podría cambiar la distribución de las mismas. Por otro lado, para las variables restantes, se imputa su valor faltante usando la media, logrando que ninguna variable contenga valores nulos.

Posterior a obtener el 100% de completitud para todas las variables, se procedió a analizar por separado las variables que son cuantitativas y las que son cualitativas, como se muestra en la Figura 4.

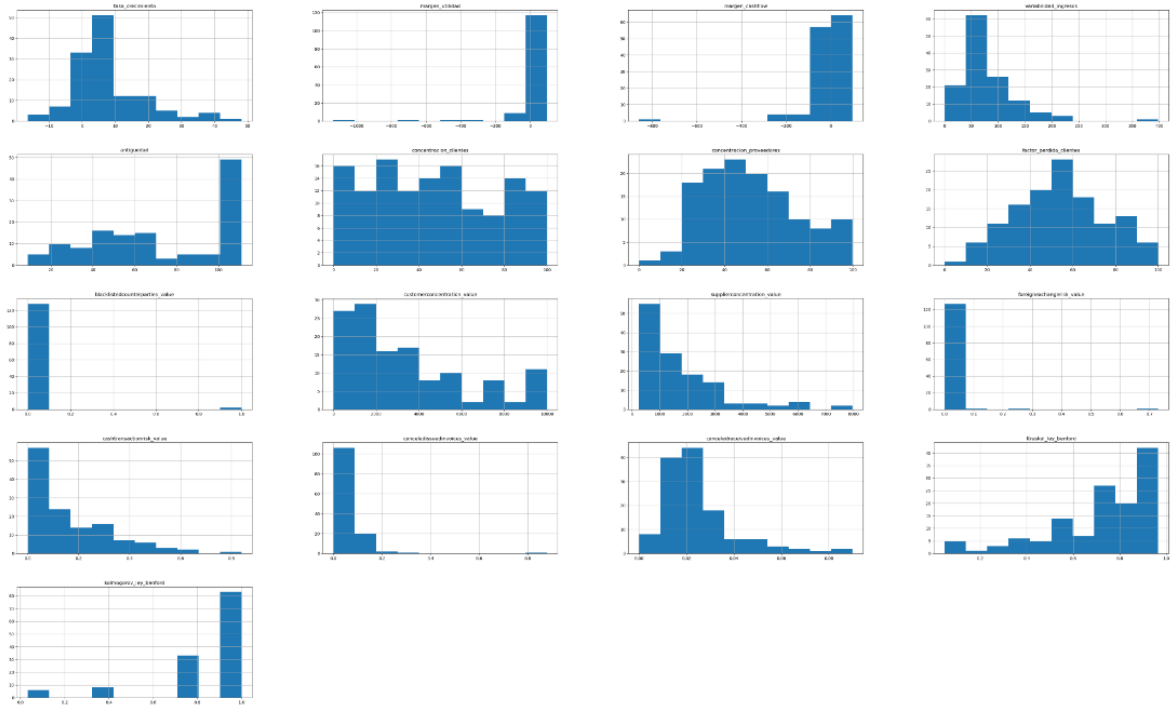

Figura 4. Distribución de las variables continuas (Puente, 2023)

Realizando un análisis sobre las variables cuantitativas, hay dos variables, “foreignexchangerisk_value” y “blacklistedcounterparties_value” que no varían, sus registros son prácticamente todos 0, por lo que son eliminadas.

También se hizo un análisis de correlaciones sobre las variables continuas.

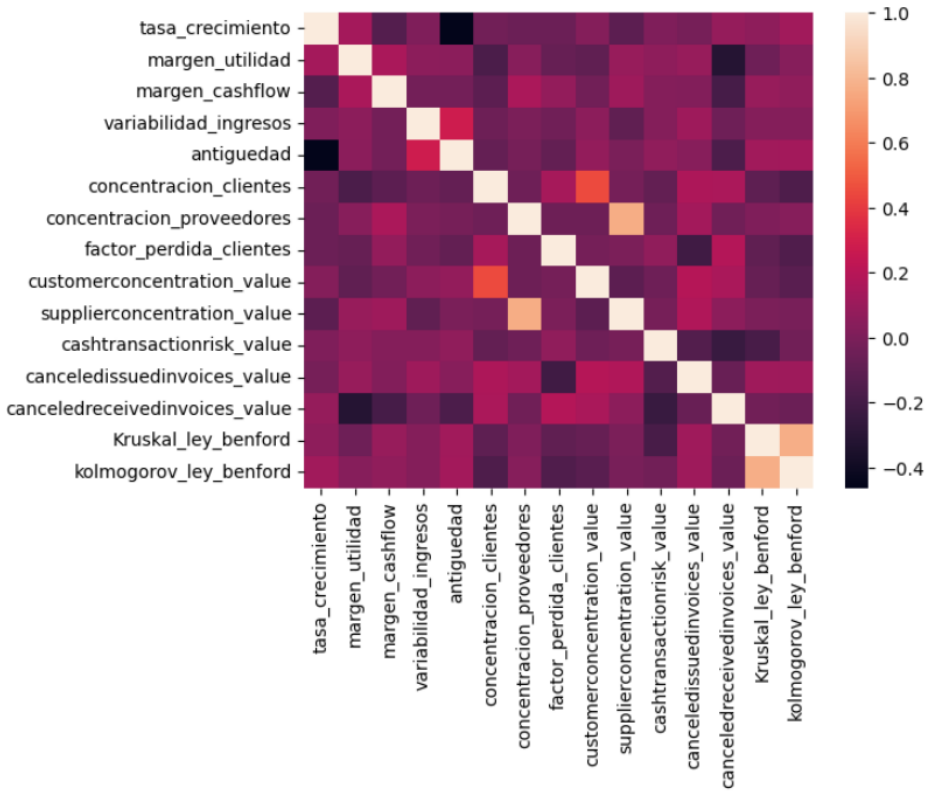

Figura 5. Correlación de variables cuantitativas (Puente, 2023)

Tras analizar las correlaciones, se detectó que:

- La variable “concentracion_proveedores” estaba altamente correlacionada con “supplierconcentration_value” por lo que esta última fue eliminada debido a que la primera es construcción por reglas de negocio y se dio prioridad a esa información.
- La variable “Kruskal_ley_benford” estaba altamente correlacionada con “kolmogorov_ley_benford” por lo que esta última fue eliminada.

Por otro lado, sólo se cuenta con una variable categórica, “taxcompliance_value_positive”, la cual, sólo contiene dos categorías “positive” y “negative” por lo que se genera una variable

dummie para poderla incluir en el modelado, donde si vale 1 es “positive” y 0 en caso contrario.

Conjunto de Entrenamiento y Validación

Resulta de gran importancia dividir el conjunto de datos original porque en entrenamiento y validación porque: el conjunto de entrenamiento se utiliza para ajustar los parámetros del modelo y aprender los patrones de los datos, mientras que el conjunto de prueba se utiliza para medir el rendimiento final del modelo después de que se hayan ajustado todos los parámetros y se hayan elegido los mejores hiperparámetros. Esto permite evaluar la capacidad del modelo para generalizar a datos nuevos y no vistos.

Por lo que se dividió en *75%* (97 registros) *entrenamiento* y *25%* (33 registros) *validación*. Para asegurar que los porcentajes establecidos fueran correctos se hizo lo siguiente:

- Se validó que la variable objetivo mantuviera su distribución de las clases en ambos conjuntos y el original.
- Se validó que las variables independientes cuantitativas mantuvieran su distribución comparando el conjunto original, el de entrenamiento y el de validación mediante la *prueba de bondad de ajuste de Kruskal-Wallis*. Tabla con los p-value de la prueba en el apéndice.
- Se validó que la variable independiente cualitativa mantuviera su distribución en los tres conjuntos: original, entrenamiento y validación.

Balanceo de Clases

Dado que la variable dependiente estaba muy desbalanceada (véase en la Figura 6), se utilizaron dos técnicas de muestreo para hacer el balanceo, con el objetivo de medir con cuál muestra lograría generalizar mejor el modelo ganador.

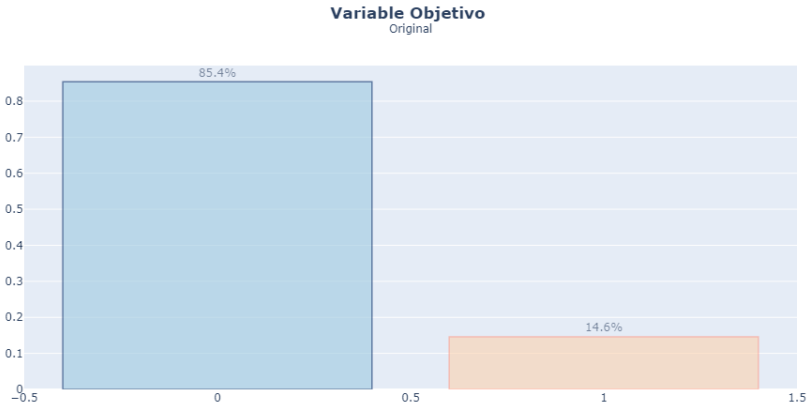

Figura 6. Distribución de las clases de la variable objetivo (Puente, 2023)

- Utilizando *undersample* se igualaron las clases conservando solamente 28 registros.
- Usando *oversample* se igualaron las clases conservando 166 registros.

Modelado

Se probaron dos metodologías: *Máquinas de Soporte Vectorial (SVM)* y *Redes Neuronales (RN)*. Para ambos casos se aplicaron todos los conjuntos de datos que se generaron (original, undersample y oversample) y se les aplicó una búsqueda de hiperparámetros obteniendo lo mostrado en el Cuadro 2.

Cuadro 2. Comparación de modelos entrenados.

Modelo	Muestra	Parámetros	Accuracy Entrenamiento	Accuracy Validación
SVM	Original	'C': 2,335,721.47, 'degree': 2, 'kernel': 'linear'	0.86	0.82
SVM	Undersample	'C': 4.83, 'degree': 2, 'kernel': 'linear'	0.82	0.58
SVM	Oversample	'C': 20.69, 'degree': 2, 'kernel': 'rbf'	1	0.75
RN	Original	'activation': 'relu', 'hidden_layer_sizes': (200, 200), 'learning_rate': 'invscaling', 'learning_rate_init': 0.001, 'solver': 'sgd'	0.33	0.30
RN	Undersample	'activation': 'relu', 'hidden_layer_sizes': (5, 5), 'learning_rate': 'invscaling', 'learning_rate_init': 0.001, 'solver': 'sgd'	0.64	0.64
RN	Oversample	'activation': 'relu', 'hidden_layer_sizes': (20,), 'learning_rate': 'adaptive', 'learning_rate_init': 0.1, 'solver': 'adam'	0.70	0.64

(Puente, 2023)

Posterior a comparar todos los modelos, de cada metodología se destacaron los que obtuvieron mejor rendimiento.

- SVM – Original: Fue el que mejor accuracy obtuvo tanto en entrenamiento como en validación, sin embargo, tras obtener otras métricas como el F1 Score (ayuda a medir el rendimiento de modelos de clasificación con clases desbalanceadas) y la matriz de confusión se observó que todo prácticamente todo lo clasifica como 0.
- RN – Oversample: Fue el mejor modelo porque sus métricas con el entrenamiento y el de validación no difieren mucho y también obtiene buen F1 score.

Mejor Modelo

Se selección el modelo de Red Neuronal con la muestra de Oversample como el mejor, debido a su rendimiento, por lo que se graficó la curva ROC con el conjunto de prueba para ver su comportamiento.

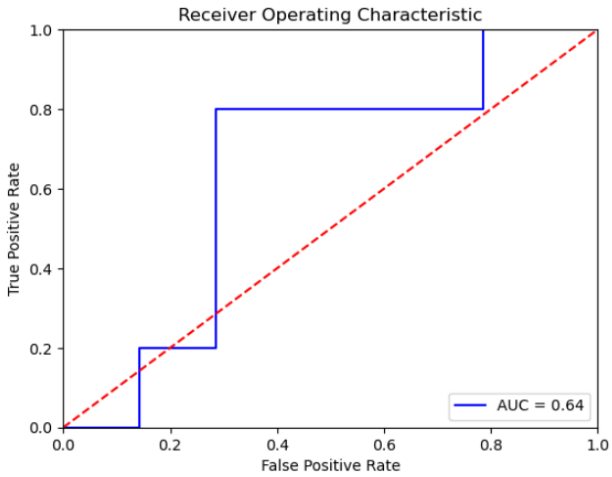

Figura 7. Curva ROC del mejor modelo (Puente, 2023)

Como se aprecia en la Figura 7, el modelo tiende a ser mejor que el azar, por lo que indica que se puede usar para hacer las clasificaciones. Sin embargo, también opté por analizar la distribución de las probabilidades de que un cliente nuevo cayera en impago diferenciando por lo que sucedió realmente.

Figura 8. Distribución de la Probabilidad de Impago en el conjunto de validación (Puente, 2023)

Tras analizar la Figura 8, es posible percatarse que la distribución de los buenos (los que no cayeron en mora) tiende a concentrarse del lado izquierdo, es decir, a tener baja probabilidad de impago, mientras que los malos (los que cayeron en mora) se concentran en los valores medios, sin embargo, existe empalme en las distribuciones e incluso hay buenos que el modelo clasifica que caerán en impago, aunque son la minoría. El escenario ideal es que ambas distribuciones estén bien separadas una de la otra, siendo los malos quienes deberían de tener alta probabilidad de impago.

Conclusiones y Recomendaciones

En conclusión, se logró ajustar con éxito un modelo de red neuronal que estimara la probabilidad de impago de los clientes, a pesar, que era un gran reto, debido a que fueron "pocos" registros y no se está utilizando información respecto del comportamiento crediticio de los clientes, solamente información resumida de la facturación emitida y recibida. Obteniendo una herramienta más robusta estadísticamente y computacionalmente que simples reglas de negocio.

Sin embargo, se recomienda que no se utilice como la única métrica en la decisión de otorgar un crédito, que su uso, de momento, vaya enfocado en que ser un indicador más que considere el equipo de analistas, debido a que por la poca información con la que se entrenó, produce predicciones con cierto margen de error. Por otro lado, se aconseja que se obtenga información del comportamiento crediticio de los clientes como el que ofrece Buró de Crédito o Círculo de Crédito y así lograr mejorar significativamente la precisión del modelo y en un futuro lograr automatizar el proceso de otorgamiento de créditos.

Apéndice

Diccionario de Variables

Debido a que se utilizó información sensible y reglas de negocio, no se puede describir con exactitud el cálculo o qué contiene cada variable, sin embargo, el nombre de las mismas es bastante descriptivo.

Cuadro 3. Variables y tipos de dato

Variable	Tipo de Dato
id_cliente	ID
par30	Binario
tasa_crecimiento	Flotante
margen_utilidad	Flotante
margen_cashflow	Flotante
variabilidad_ingresos	Flotante
antiguedad	Entero
concentracion_clientes	Flotante
concentracion_proveedores	Flotante
factor_perdida_clientes	Flotante
taxcompliance_value	Cadena
blacklistestatus_value	Flotante
blacklistedcounterparties_value	Entero
intercompanytransactions_value	Flotante
customerconcentration_value	Entero
supplierconcentration_value	Entero
foreignexchangerisk_value	Flotante
cashtransactionrisk_value	Flotante
accountinginsolvency_value	Flotante
canceledissuedinvoices_value	Flotante
canceledreceivedinvoices_value	Flotante
Kruskal_p_value	Flotante
kolmogorov_p_value	Flotante

(Puente, 2023)

Resultados de Prueba de Kruskal-Wallis

Cuadro 4. Resultados de aplicar la prueba de Kruskal-Wallis a variables cuantitativas

Variable	P-value
tasa_crecimiento	0.8992
margen_utilidad	0.1414
margen_cashflow	0.4724
variabilidad_ingresos	0.9744
antigüedad	0.7503
concentracion_clientes	0.5516
concentracion_proveedores	0.4002
factor_perdida_clientes	0.5371
customerconcentration_value	0.6051
cashtransactionrisk_value	0.9994
canceledissuedinvoices_value	0.6594
canceledreceivedinvoices_value	0.8048
Kruskal_ley_benford	0.4822

(Puente, 2023)

Referencias

AWS. (2023). *¿Qué es una red neuronal?* Obtenido de AWS: https://aws.amazon.com/es/what-is/neural-network/#:~:text=Una%20red%20neuronal%20es%20un,lo%20hace%20el%20cerebro%20humano.

Badr, W. (22 de 02 de 2019). *Having an Imbalanced Dataset? Here Is How You Can Fix It.* Obtenido de Medium: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb

Conover, W. J. (1999). *Practical nonparametric statistics (3rd ed.).* John Wiley & Sons.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* The MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.).* Springer.

Puente, A. (2023). Proyecto_AndrePuente.ipynb.