MODELO DE APRENDIZAJE AUTOMÁTICO PARA RIESGO CREDITICIO DE MICROEMPRESARIOS REGIONALES SEGÚN PERFIL SOCIOECONÓMICO

AUTOMATIC LEARNING MODEL FOR CREDIT RISK OF REGIONAL MICRO ENTREPRENEURS ACCORDING TO SOCIOECONOMIC PROFILE

Carlos Mario Betancur Londoño, cmbetancur@libertadores.edu.co
José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co

RESUMEN

La cartera de crédito es fundamental en una entidad financiera, por ello ante cada crédito entregado, la esperanza es recuperarla en tiempos pactados con el cliente, aún así, es latente el riesgo de no pago durante la vigencia de la obligación. La propuesta de un modelo de predicción con diferentes técnicas que defina la probabilidad de default, puede ayudar a definir las posibles causas socioeconómicas que implican riesgo de impago. Se tomó la consecución de los defaults causados en el instante con el objetivo de identificar clientes que podrían incurrir en estado de mora y riesgo de no pago. La modelación se hizo con el fin de mitigar o filtrar los usuarios a los cuales se les otorga el crédito y nos ayuda a definir cómo puede ser catalogado como habiente potencial de default, esto, determinado por los perfiles que nos proveen los más de 39 mil individuos que conforman la base de datos. El nicho de mercado al cual se dirige la institución, está conformado por usuarios con alcance económico limitado para iniciar su empresa o microempresarios que requieren capital de trabajo para su negocio en curso, todos ellos con un interés en común, crear empresa y salir adelante con su idea, sin importar niveles académicos, músculo financiero o residencia urbana o rural. Es menester un concepto sólido del proyecto y su puesta en marcha. Es fundamental tener claro el nicho de mercado al cual se dirige la institución y por ello es relevante considerar cuál es el perfil que lo conforma. Los modelos expuestos en este proyecto tienen fundamentos de apoyo para el área de estudios de crédito o central de evaluación financiera. El procedimiento de modelado se llevó a cabo con métodos de aprendizaje máquina supervisados como regresión logística, random forest y gradient boosting. Tres opciones de las cuales se escogió el random forest como la mejor, según sus

métricas. Se hizo el comparativo con la metodología actual de evaluación crediticia y se determinó las implicaciones en caso de ser implementado.

Palabras clave: Machine learning, aprendizaje automático supervisado, riesgo crediticio, default, estudio de crédito.

ABSTRACT

The credit portfolio is fundamental in a financial entity, therefore, before each credit delivered, the hope is to recover it in times agreed with the client, even so, the risk of non-payment during the term of the obligation is latent. The proposal of a prediction model with different techniques that defines the probability of default, can help define the possible socioeconomic causes that imply risk of default. The achievement of the defaults caused at the moment was taken with the objective of identifying clients that could incur in a state of default and risk of non-payment. The modeling was done in order to mitigate or filter the users to whom the credit is granted and helps us define how they can be classified as a potential default holder, this, determined by the profiles provided by the more than 39 thousand individuals that make up the database. The market niche to which the institution is directed is made up of users with limited economic scope to start their business or micro-entrepreneurs who require working capital for their ongoing business, all of them with a common interest, to create a business and get ahead with your idea, regardless of academic levels, financial muscle or urban or rural residence. A solid concept of the project and its implementation is necessary. It is essential to be clear about the market niche to which the institution is directed, and for this reason it is important to consider what its profile is. The models exposed in this project have foundations of support for the area of credit studies or central financial evaluation. The modeling procedure was carried out with supervised machine learning methods such as logistic regression, random forest and gradient boosting. Three options of which the random forest was chosen as the best, according to its metrics. The comparison was made with the current credit evaluation methodology and the implications were determined in case of being implemented.

Keywords: Machine learning, supervised machine learning, credit risk, default, credit study.

INTRODUCCIÓN

El riesgo es una situación que permanece latente en cualquier institución, esta es definida como el conjunto de posibilidades de que se produzca un contratiempo o perjuicio. Asimismo, la ocurrencia tiene una consecuencia o resultado definido por el tamaño y éste a su vez será la magnitud del riesgo. Existen muchas técnicas para su evaluación y forma de mitigarlo.

Entre los múltiples riesgos que pueden acarrear las empresas en el desarrollo de su operación, existen los denominados riesgos financieros y se manifiestan en las diferentes actividades que desempeña, todos y cada uno de ellos con posibles hechos consecuentes como resultados negativos o choques inesperados debido a diferentes causas como toma de decisiones sin bases bien definidas, endeudamientos no estudiados, cambios en el mercado, especulaciones, etcétera. Los riesgos financieros están conformados principalmente por los siguientes (Banco de la República).

a. Riesgo de Mercado, definido como la posibilidad de obtener pérdidas en los valores de posición por cambios en las variables que inciden en la valoración de productos, servicios o activos financieros.
b. Riesgo de Liquidez, cuya implicación es no tener alcance para soportar el desarrollo de su actividad. Por lo que sugiere acudir a la venta de sus activos castigando su valor en el mercado.
c. Riesgo Operacional, con una definición propuesta por el banco internacional de Basilea, es la situación en que se puede provocar pérdidas como resultado de errores humanos, procesos inadecuados o defectuosos, errores en los sistemas y por acontecimientos externos.
d. Por último, Riesgo Crediticio, cuya trascendencia en una empresa es fundamental. Está definido como la probabilidad de que una de las partes del contrato incumpla sus obligaciones por insolvencia, enfermedad u otras causas que impliquen incapacidad de pago produciendo pérdida financiera a la otra parte (Sagner T, 2012). En el sector financiero existe una definición más segmentada propuesta por el comité de Supervisión Bancaria de Basilea, es la posibilidad de que un prestatario bancario o una contraparte no cumpla con sus obligaciones de acuerdo a los términos acordados (Basel Committee on Banking Supervision, 1999, p. 1)

En el presente trabajo se considera el último tipo de riesgo como base para su desarrollo considerando el campo de aprendizaje automático.

El proyecto que se desarrollará, pone en común el machine learning como herramienta de la inteligencia artificial para la transformación de datos y cálculos de gran magnitud, con el estudio del riesgo crediticio en instituciones financieras como cooperativas y bancos. La propuesta conformará una herramienta que servirá de apoyo a las políticas de crédito, cuantificando las probabilidades de incumplimiento o default y definir una mayor precisión sobre la clasificación de clientes potenciales del otorgamiento del crédito. Para este enfoque se acude a los datos financieros con un número de variables cuantitativas y

cualitativas. Variables demográficas, de situación financiera y de mercado según el destino del crédito.

Entre los posibles modelos que se desarrollarán, existen la regresión logística y otros como el random forest o árboles de decisión y el gradient boosting, con la espera de unos resultados producidos que sean fáciles de interpretar (Grau, 2020). Esta esperanza está dada tanto por el modelo investigado como por la calidad de datos de partida. Con un modelo errado o una calidad de datos poco relevante, la efectividad sería de baja confiabilidad. De esta misma forma, el entendimiento de los datos debe ser propicia desde un principio, lo que conlleva a una selección de información clara, entendible y descriptiva para lograr obtener una muestra que no nos lleve a posibles errores en sí misma.

Para el avance del proyecto, existe una claridad acerca de la obtención de la base por la propia protección de datos, ya que estos pueden contener uno sensibles pero que son verdaderamente importantes para el entrenamiento del modelo. Adicional, las instituciones financieras se comportan con recelo considerando que estos datos son la propia fuente de sus modelos. Esta puede ser una observación para indicar que no sería sencillo contar con los datos amplios y obtener un modelo predictivo de eficacia óptima.

La descripción hecha en los anteriores párrafos nos lleva a pensar en el objetivo que tendrá el presente proyecto, para ello, las estimación y precisiones del modelo de riesgo de crédito, se hará en consideración de una cartera de microempresarios cuya financiación obtenida fue para su negocio. En el desarrollo del proceso se irá determinando cuáles variables son verdaderamente relevantes para la evaluación. Para fines netamente académicos, se procede con una metodología bajo el aprendizaje supervisado.

REFERENTES TEÓRICOS

Conservando los objetivos propuestos del presente proyecto, se han presentado diversos documentos que describen el uso y aplicación de machine learning y los temas del riesgo de crédito. Una publicación muy amplia en su descripción ha sido “Propuesta de Modelo para evaluación de Riesgo de Crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito LA” (Cuenca et, 2019). Manifiesta que la regresión logística se desempeña como mejor modelo dado su rendimiento y sus métricas para la predicción del riesgo, su modelo presentó un accuracy de 0.6634, sensibilidad de 0.6448, especificidad 0.6821 y precisión de 0.6698. Con herramientas como la curva de características operativas receptoras (ROC), su área bajo la curva (AUC) y favorece el descarte de modelaciones econométricas, que han sido usados convencionalmente. Asimismo, hace reconocimiento

a las posibles dificultades que puede presentar la modelación, entre ellas se encuentran la calidad de los datos, el filtro de las variables y aplicación de algoritmos.

La regresión logística cuenta con ciertas limitaciones a la hora de aplicar el modelo, afectaciones subyacentes como la apropiada vinculación de las variables, que podrían afectarse por multicolinealidad. Para ellos surgen nuevas herramientas como el random forest, cuyas métricas lo catapultan como superior por sus estimaciones. (Kruppa et al., 2013). Una revisión sistemática de la literatura, lleva a definir que no existe consenso alguno en cuanto al rendimiento de los modelos estadísticos y machine learning para determinar la calificación crediticia. (Dastile et al., 2020). Esto, porque pueden presentar dificultades en la explicación de las predicciones y datos desequilibrados.

METODOLOGÍA

Origen de los datos

Este proyecto tuvo como precedente, la metodología de estudio de crédito que se lleva a cabo en la empresa Microempresas de Colombia Cooperativa A C, a la hora de presentar una solicitud crediticia. Para ello se usó la base de datos anonimizada y de acceso privado para la institución. Ésta cuenta con $n = 39789$ individuos, donde cada uno representa una obligación única de cartera. La selección de la base se hizo considerando diferentes periodos del año 2021. Su escogencia fue de un mes aleatorio entre los 12 posibles del año. Inicialmente se contó con más de 100 columnas que categorizaban al usuario, pero fueron disminuidas con una evaluación hecha en compañía de las áreas de cartera y riesgo porque eran parte de la parametrización del sistema. Se concluyó que sólo 18 de éstas (9 cuantitativas y 9 categóricas) aportaban a la descripción del perfil según su situación social, demográfica, económica, tipo de negocio y estado del crédito.

Preparación de datos y variables

El desarrollo del proyecto se hizo con Python basado en la versión 3.10.5 en la interfaz Google colab. Para el modelo basado en reglas se utilizó el paquete SweetViz v.2.1.4 y el planteamiento del modelo basado en machine learning tuvo como herramientas los paquetes PyCaret (https://pycaret.org/) y Scikit Learn (https://scikit-learn.org/stable/). El código concluido se halla en el siguiente enlace.
https://colab.research.google.com/drive/1JoHnDBgE-bzAMpC-MNaNztztLoejHHJe?usp=sharing

El tratamiento y preparación de datos se llevaron a cabo con la imputación de datos ausentes, incorporando la moda en las variables categóricas y la media en las numéricas. No existieron causas suficientes para el descarte de individuos por datos vacíos, así se dio

uso de la totalidad de los registros. Para cada uno de ellos se usó la edad de cartera (Días_Vencido) como variable objetivo y se dicotomiza en una nueva variable llamada Default, dónde (1) es la mayor o igual a 180 días ($edad \geq 180\ días$) y (0) la menor ($edad < 180\ días$).

Para las variables numéricas se usó la transformación logarítmica base 10 con el fin de mejorar la escala y representación de la medida y adecuar las cifras para el entrenamiento del modelo predictivo. (Tabla 1). Y se realizó un análisis descriptivo exploratorio univariado y relacional entre variables.

Tabla 1. Definición y transformación de variables

Variable	Descripción	Unidad	Tipología	Transformación
Sucursal	Oficina donde se gestiona la colocación	Sin unidad	Categórica	
Región	Región del país a la cual pertenece la sucursal	Sin unidad	Categórica	
Est_Civil	Estado civil del individuo al momento de realizar el crédito	Sin unidad	Categórica	
Sexo	Género femenino o masculino del individuo	Sin unidad	Categórica	M:0 H:1
Nivel_Estudio	Nivel académico que posee el individuo	Sin unidad	Categórica	
Edad	Número de años vividos del solicitante	años	Numérica	Log10
Estrato	Estrato social al cual pertenece el individuo o su empresa	Sin unidad	Categórica	
Tipo_Vivienda	Define el tipo de vivienda donde reside	Sin unidad	Categórica	
Ingresos	Nivel de ingresos que genera en su actividad principal o empresa	Pesos colombianos	Numérica	Log10
Otro_Ingreso	Nivel de ingresos no soportados o que no son de la actividad principal	Pesos colombianos	Numérica	Log10
Egreso	Nivel de salida de dinero que posee como persona	Pesos colombianos	Numérica	Log10
Tipo_Crédito	Naturaleza de la dauda y a la cual se destina el dinero adquirido	Sin unidad	Categórica	
Monto	Valor por el cual se le otorga el crédito	Pesos colombianos	Numérica	Log10
Cuotas	Número de cuotas pactadas para cubrir la deuda	Cuotas	Numérica	Log10
Vlr_Cuota	Valor de la cuota pactada para cubrir la deuda en un número de cuotas	Pesos colombianos	Numérica	Log10
Actividad_Comercial	Sector comercial en el cual desempeña	Sin unidad	Categórica	
Días_Vencido	Número de días transcurridos desde la última cuota pagada	Pesos colombianos	Numérica	
Saldo_Actual	Saldo vigente en el periodo evaluado	Pesos colombianos	Numérica	Log10

Integración de datos y modelación

Modelo basado en reglas

Una vez hecho el procedimiento anterior, se planteó un modelo basado en reglas donde se esperaba que las regiones como bajo cauca, córdoba, occidente y la edad transformada menor a 25 años, obtuvieran resultados como posibles defaults. Para el caso de las regiones, se esperaba tal situación por las dificultades de orden público en el bajo cauca y córdoba y el difícil acceso a comunicaciones desde las zonas rurales en la zona de occidente. Antes de iniciar con el modelo base, la cartera poseía $n = 37960$ individuos sin default y $n = 1829$ con riesgo de impago, lo que representaba el 95% y 5% respectivamente. Para el balanceo de clase en este modelo, se aplicó la metodología de submuestreo, permitiendo la selección aleatoria de $n = 1829$ individuos sin riesgo de impago para igualarlo en cantidad al grupo que sí estaba definido como default. Así, se obtuvo la variable cartera3, variable que describió una cartera con $m = 3658$ individuos y balanceada 50/50 con $n_1 = n_2 = 1829$.

Posteriormente, la base de cartera fue dividida en dos grupos definidos para el entrenamiento y testeo del modelo. El primer grupo de entrenamiento, cubriendo un 80% ($n = 2926$) y el segundo para testeo o prueba con el 20% restante ($n = 732$).

Modelo Machine Learning (ML)

En el planteamiento del modelo machine learning, el problema de desbalanceo de clase fue tratado con la librería Así, la creación del modelo reportó una serie de posibles métodos de clasificación, con la variable objetivo default como salida. Los modelos escogidos fueron los planteados como objeto de estudio en este proyecto. Random forest, regresión logística y gradient boosting. Así, se procedió con el análisis de la matriz de confusión, las métricas de desempeño como Accuracy (exactitud) o proporción de predicciones que el modelo clasificó correctamente, Recall (sensibilidad) como proporción de positivos reales identificados correctamente, Precisión o proporción de instancias relevantes entre las recuperadas, f1 score de robustez y precisión dada la media armónica de recuperación y las respectivas curvas ROC y su AUC. Cada una de las métricas definidas formalmente como sigue.

Ecuación 1. Fórmula Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ecuación 2. Fórmula Precisión

$$Precision = \frac{TP}{TP + FP}$$

Ecuación 3. Fórmula Recall

$$Recall = \frac{TP}{TP + FN}$$

Ecuación 4. Fórmula F1Score

$$F1_{Score} = 2 * \frac{Accuracy * Recall}{Accuracy + Recall}$$

Donde:

$TP = Verdaderos\ Positivos\ (True\ Positives)$
$TN = Verdaderos\ Negativos\ (True\ Negatives)$
$FP = Falsos\ Positivos\ (False\ Positives)$
$FN = Falsos\ Negativos\ (False\ Negatives)$

ROC-AUC

El último indicador del desempeño de los distintos métodos aplicados para la clasificación de clientes morosos y no morosos es la Curva Característica Operativa del Receptor, ROC (por sus siglas en inglés), el cual es un gráfico que muestra la capacidad de un sistema de clasificación binario para realizar un diagnóstico. Inicialmente el ROC se utilizó en los sistemas de detección de señales, por lo que sus primeras aplicaciones fueron en el campo militar.

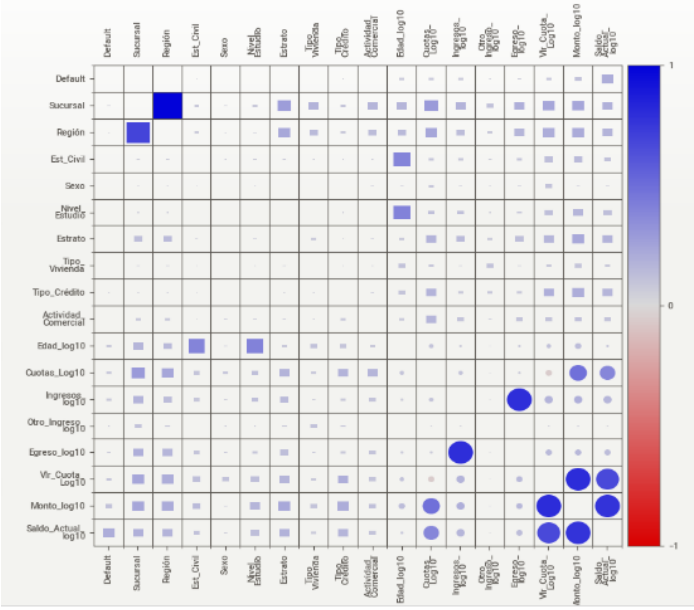
Desde el punto de vista interpretativo, el ROC es una descripción del efecto del umbral de detección de la clasificación binaria, mostrando todas las posibles combinaciones de las frecuencias relativas de clasificaciones correctas e incorrectas (Metz, 1978); es decir, la tasa de acierto y la tasa de falsas alarmas (Fawcett, 2006).

Los dos anteriores párrafos descritos como definición de la curva y su área. (Ossa y Jaramillo, 2020).

RESULTADOS

Análisis descriptivo

Figura 1. Los cuadrados son asociaciones categóricas (coeficiente de incertidumbre y relación de correlación) de 0 a 1. El coeficiente de incertidumbre es asimétrico (es decir, los valores de etiqueta de fila indican cuánta información proporcionan a cada etiqueta en la parte superior). Los círculos son las correlaciones numéricas simétricas (de Pearson) de -1 a 1. La diagonal trivial se deja en blanco intencionalmente para mayor claridad.



La variable objetivo no presenta relación significativa con ninguna de las variables categóricas o numéricas, la mayor relación que puede existir, es con el valor del saldo actual. Las demás variables como la sucursal, tiene relación con la región, como es de esperarse porque los grupos de cada región lo constituyen las sucursales. Existen otras relaciones representativas como la de ingresos con egresos, el valor de la cuota con el monto solicitado y ésta a la vez con el saldo del instante. Unas muy leves pero que se pueden resaltar, son el número de cuotas y el monto desembolsado o el saldo actual de la deuda. Para las demás caracterizaciones, no son relevantes las asociaciones, por lo que se pueden omitir del análisis.

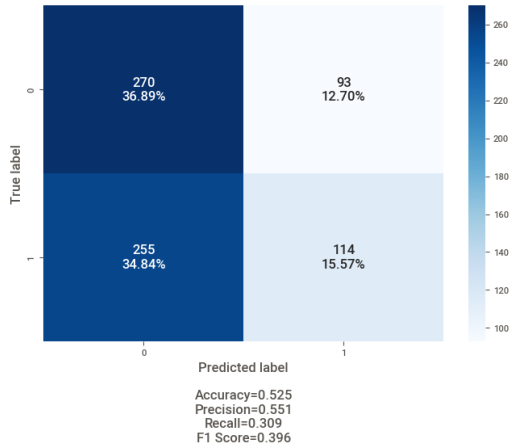
La edad promedio de los individuos estudiados es cerca de los 44 años, edad muy próxima a la del 50% de ellos. El usuario menor ronda los 18 años, mientras el mayor supera por poco los 87. Los ingresos de la población están en promedio de $1.99 millones de pesos y van desde usuarios que no tienen hasta otros cuyo ingreso llega a $83 millones y un 50% de ellos llegan a tener no más de $1.5 millones. El 75% de los usuarios reportan no tener otros ingresos. Para los egresos se evidencia que el promedio puede llegar a $1.04 millones y la mitad de los asociados generan egresos no mayores a los $0.76 millones. El máximo de egreso reportado llega a los $6.4 millones. Siendo así, los montos solicitados como desembolso de crédito son el promedio $5.2 millones por obligación y la mitad de la población de cartera solicita hasta $3.5 millones, los demás desembolsar crédito cuyo capital alcanza los $100 millones. La edad de la cartera, como factor muy importante en este proyecto, tiene en promedio un poco menos de 21 días y un 75% alcanzan mora de 13 días, aun así, tenemos que un 25% restante alcanzan edad de incumplimiento hasta los 479 días. Para los desembolsos ejecutados a la fecha se tiene que, en promedio, los saldos adeudados llegan a $3.5 millones por crédito vigente. El 25% tiene vigente hasta $1.1 millones, un 50% hasta $2.15 millones y el 75% hasta $3.95 millones unos máximos de hasta $94 millones.

La cartera ha sido desembolsada en 30 oficinas. La oficina que más desembolsos generó fue Villanueva con 5171 créditos y hace parte de la región Valle de Aburrá que también se describe como la que mayor número de créditos ha desembolsado con 9741. De la totalidad de créditos, 15493 se han otorgado a usuarios cuyo estado civil es unión libre y el género femenino es quien más oportunidad de endeudamiento tuvo participando con 21186. También se puede identificar que, de todos los usuarios, 16308 alcanza niveles académicos de la secundaria y así mismo 17121 hacen parte del estrato 2. La casa propia es un factor verdaderamente influyente a la hora que han solicitado el crédito, 35943 posee este tipo de vivienda. La actividad principal del solicitante, como se esperaba, ha sido para

microempresarios y participan con 38757 de los créditos evaluados y 35830 han sido para microcrédito.

Considerando lo observado en el análisis descriptivo, el modelo heurístico consistió en el planteamiento de las condiciones descritas en la integración de datos, dónde las regiones mencionadas y la edad, nos permitirían determinar qué usuarios podrían ser potenciales default. Midiendo su desempeño se obtuvieron las métricas respectivas. El Accuracy de 52.46% (Figura 2). Métrica que nos ayudará a realizar la comparativa contra los demás modelos ML y lograr definir su aplicabilidad (Tabla 2).

Figura 2. Matriz de confusión modelo basado en reglas. Datos balanceados con submuestreo.



A continuación, en la tabla 2 se evidencian los cinco mejores modelos ML, en su mayoría clasificadores y con la métrica de comparación (Accuracy) muy similares entre sí y superiores al 89%. No obstante, en consideración a lo planteado desde el inicio, el enfoque está definido para el modelo de bosques aleatorios (random forest), que se encuentra entre la lista relacionada y de la cual no acoge las otras técnicas dichas, regresión logística y aumento de gradiente (gradient boosting).

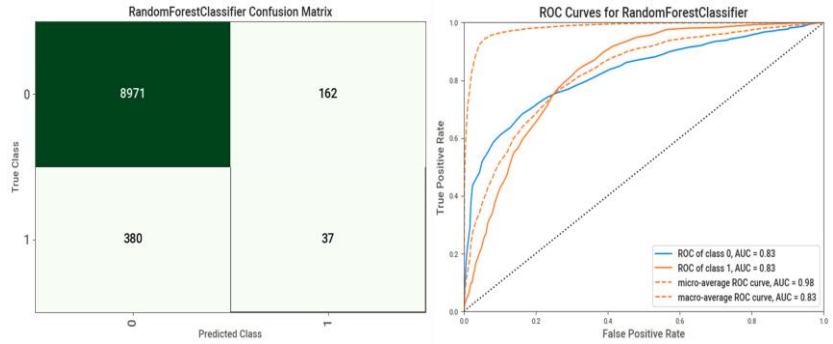
Tabla 2. Desempeño promedio de los mejores modelos ML.

	Model	Accuracy	AUC	Recall	Prec.	F1
MBR	Modelo Basado en Reglas	0,5246	0,5264	0,3089	0,5507	0,3958
dummy	Dummy Classifier	0,9539	0,5000	0,0000	0,0000	0,0000
rf	Random Forest Classifier	0,9428	0,8117	0,0972	0,2231	0,1353
et	Extra Trees Classifier	0,9374	0,7187	0,0758	0,1499	0,1005
lightgbm	Light Gradient Boosting Machine	0,9323	0,8295	0,1546	0,1991	0,1735
dt	Decision Tree Classifier	0,8924	0,5955	0,2685	0,1445	0,1875

Bien es cierto que la técnica random forest no presenta la mayor exactitud promedio entre los modelos, aun así, el resto de métricas presentan mejores indicadores que el "dummy", por ello sigue siendo viable considerarlo como el mejor modelo. Siendo así, procedemos a profundizar en el ROC y la matriz de confusión (Figura 3) para evaluar su viabilidad.

La técnica ha asignado 37 default con acierto positivo, por lo que se evitaría errar en este número de créditos otorgados. Lo contrario ocurre con 162 créditos que se negarían porque no siendo realmente default, la técnica los definió como riesgosos ante el impago de la deuda. Por otro lado, teniendo conocimiento que 380 usuarios ya son un riesgo por el impago causado, el modelo los determina como no riesgosos. La gran mayoría de las solicitudes estudiadas, 9550 en total, el algoritmo determinó que 8971 son verdaderamente poco riesgosos, siendo reales en lo ya causado. Así, 9008 clasificaciones fueron definidas correctamente y 542 se catalogan como errores en la predicción. Se cuantificará las implicaciones monetarias que los errores pueden representar con base en el promedio de los montos solicitados según el análisis descriptivo.

Figura 3. Matriz de confusión y curva ROC del modelo random forest.



Comentado [BLCM1]: No me fue posible el cambio de lenguaje en el gráfico.

DISCUSIÓN

El riesgo crediticio es una situación a la que cualquier institución financiera está sometida. Existen diferentes métodos para mitigarlo y entre ellos está el estudio del perfil socio-económico del solicitante. Por ello, el presente proyecto se desarrolla en aras de pronosticar posibles situaciones de impago y crear prospectos de los usuarios mediante su estudio crediticio.

Con la información que nos proporcionaron los miles de individuos tratados en el desarrollo de este trabajo, se usó la herramienta de aprendizaje automático supervisado. Como respuesta, obtuvimos que el mejor modelo para tomar decisiones de otorgar o no el acceso

crediticio fue el método basado en clasificación random forest. Los modelos planteados inicialmente como era la regresión logística y gradient boosting, no presentaron métricas a considerar. De esta manera, siguen siendo modelos que representan mejor los datos en comparación al modelo base. Las diferentes métricas nos llevan a considerar que los modelos de aprendizaje automático supervisado ofrecen mejores instancias en la evaluación crediticia.

Las variables que mayor importancia presentaron en el estudio fueron los saldos actuales, el número de cuotas, el monto solicitado y el valor de la cuota, todas estas variables relacionadas directamente con las características del crédito y no con la situación socio-económica del asociado (Anexo 5). Seguidas a estas variables, caracterizaciones como la edad, género, ingresos y egresos muestran un grado de importancia ubicándose entre las ocho más importantes.

En términos de la modelación en este estudio, y como se definió en los referentes teóricos, se debía esperar que, como en otros estudios, los basados en árboles de decisión son una opción de mayor predicción considerando una cartera balanceada por la librería.

Los resultados de este estudio, nos llevan a evaluar y analizar las implicaciones que se tendría a la hora de implementarlo. El modelo reporta que existen 162 clientes potenciales a riesgo de impago, cuando en realidad no lo son. De considerarlos como nuevos solicitantes, la institución negaría su desembolso y conlleva a no desembolsar, tomando como base el promedio por cada crédito de $5.2 millones, cerca de $842 millones ($162 * 5.2\ millones$). Esto desde el escenario averso al riesgo. Asimismo, el modelo manifiesta que 380 usuarios serían libres de riesgo y podríamos desembolsar una cifra que alcanza los $2000 millones ($380 * 5.2\ millones$), cifra que, según la realidad, sí sería un riesgo de default. De todos los pronósticos, se puede desembolsar 9351 créditos equivalentes a $48.6 miles de millones ($9351 * 5.2\ millones$), con posibilidad de recuperar un capital valorado en $46.65 miles de millones ($8971 * 5.2\ millones$). Default que asciende a los cerca de $2000 millones.

En la actualidad presenta un default de $2.7 miles de millones con la cartera total de 39789 asociados, por lo que se puede inferir que no es viable la aplicación del modelo. Condicionalmente, de aplicarse en una muestra de 9550 individuos, se estaría otorgando con confianza 8971 créditos equivalentes a $46.65 miles de millones, se dejaría de desembolsar $842 millones de 162 falsos positivos y 37 créditos por $192.4 millones, por último, se entregarían 380 desembolsos equivalentes a $2000 millones, esto conlleva a una pérdida potencial por riesgo de impago y de oportunidad por rechazo, que sumando un valor aproximado de $2850 millones.

Con lo anterior, el proceso convencional, en comparación con el modelo ML, permite mejores formas de estudio de crédito. Es de considerar que la base tomada para este

proyecto no contó los créditos rechazados, sólo con la cartera existente y que hace parte del activo de la institución. Ampliar la base de datos con las solicitudes de crédito descartadas y mejorando las caracterizaciones del posible default, el aprendizaje automático supervisado puede ser un método de evaluación con mejores métricas y mejores impactos institucionales no sólo en el enfoque del crédito sino en los tiempos de estudio y respuesta.

CONCLUSIONES

El modelo de clasificación random forest, está dado como el mejor para la estimación del riesgo de crédito a microempresarios antioqueños. Al compararse con un modelo base y los demás modelos supervisados, éste representó mejores métricas y nos llevó a definirlo como superior, incluso, a la regresión logística.

Con las métricas de este proyecto, no se debe juzgar el modelo de aprendizaje automático respecto a la implementación, sólo se puede hacer la comparación con el base.

Como institución aversa al riesgo, no es viable la implementación del modelo a la forma del negocio tomando, como tabla de hechos, la información con la que se realizó este estudio. Los resultados y discusiones reflejan que su implementación puede significar pérdida en la cobertura del sector crediticio e implicaciones en los ingresos.

Mejorando la información para el entrenamiento de los modelos y subiendo un poco los niveles de riesgo asumibles, el modelo de aprendizaje automático puede ofrecer grandes ventajas desde que llega la solicitud hasta la ejecución del desembolso.

Con el modelo creado en este estudio, dada su implementación, se tendría un impacto relevante para sus ingresos, en cálculos dichos en la sección de la discusión, se dejaría de percibir los ingresos financieros que generarían los cerca de $2850 millones, ya sea porque el default ascendería a los $2000 millones o porque se dejarían de otorgar $850 millones, asimismo, la pérdida de capital a la que se estaría exponiendo la institución con el default.

El desarrollo de un modelo de aprendizaje automático puede tomar días, un tiempo que, una vez invertido, resumiría los tiempos de respuesta y sintetiza todos los tiempos en un mínimo para su estudio y su desembolso. Hecho que sería un atractivo para la institución debido a que las modalidades de estudio para los créditos a microempresarios toman días dada la intervención de humanos en el flujo de operatividad.

El conocimiento del negocio, de cuál es la dinámica de los datos, cuáles son los criterios de evaluación y los impactos de los errores en el estudio humano del crédito, son factores relevantes para la modelación y posibles mejoras para la capacidad de predicción. Alguna de las formas de implementar el aporte humano al modelo es, en un principio, la limpieza, imputación y consecución de variables. Todo esto como correcto tratamiento de datos. Asimismo, mientras se alimentan las bases de datos, se debe tener supervisión de la captura de datos.

REFERENCIAS BIBLIOGRÁFICAS

Banco de la República, Colombia. (2000). Riesgo de crédito - Informe especial de Estabilidad Financiera - Primer semestre 2021. Bogotá, https://www.banrep.gov.co/

Sagner T, A. (2012). EL INFLUJO DE CARTERA VENCIDA COMO MEDIDA DE RIESGO DE CRÉDITO: ANALISIS Y APLICACION AL CASO DE CHILE. Revista de análisis económico, 27-53.

Basel Committee on Banking Supervision (1999). Principles for the Management of Credit Risk - final document. Basel Committee on Banking Supervision, 2000(July 1999), 4.

Grau, J. (2020). MACHINE LEARNING Y RIESGO DE CRÉDITO. Universidad Pontificia COMILLAS. 44

Ltda, Merced & Cuenca, & Gonzalo, Cela & Cuenca, Juan. (2019). Propuesta de modelo de machine learning para la evaluación de riesgo de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito La.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125–5131. https://doi.org/10.1016/j.eswa.2013.03.019

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing Journal, 91, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861874. https://doi.org/10.1016/j.patrec.2005.10.010

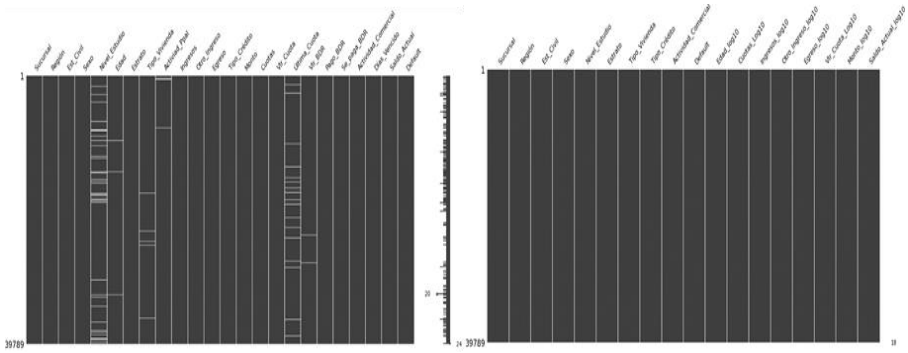
Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in Nuclear Medicine, 8(4), 283298. https://doi.org/10.1016/S0001-2998(78)80014-2

Ossa, W., Jaramillo V. Machine Learning para la estimación del riesgo de crédito en una cartera de consumo, 39. https://repository.eafit.edu.co/bitstream/handle/10784/29589/Wbeimar_OssaGiraldo_Veronica_JaramilloMarin_2021.pdf?sequence=2&isAllowed=y

Ruiz, L. (2021). FACTORES DE PREDICCIÓN DE LA APARICIÓN DE PERSONAS MAYORES REPORTADAS COMO DESAPARECIDAS, A PARTIR DE MODELOS DEAPRENDIZAJE AUTOMÁTICO SUPERVISADO. https://drive.google.com/file/d/1xranPtzEtYW8RzECP0Rw_bz9C9RsmIOY/view?usp=sharing

ANEXOS

Anexo 1. Datos ausentes tabla original e imputada.



Anexo 2. Tabla anterior y posterior a la imputación.

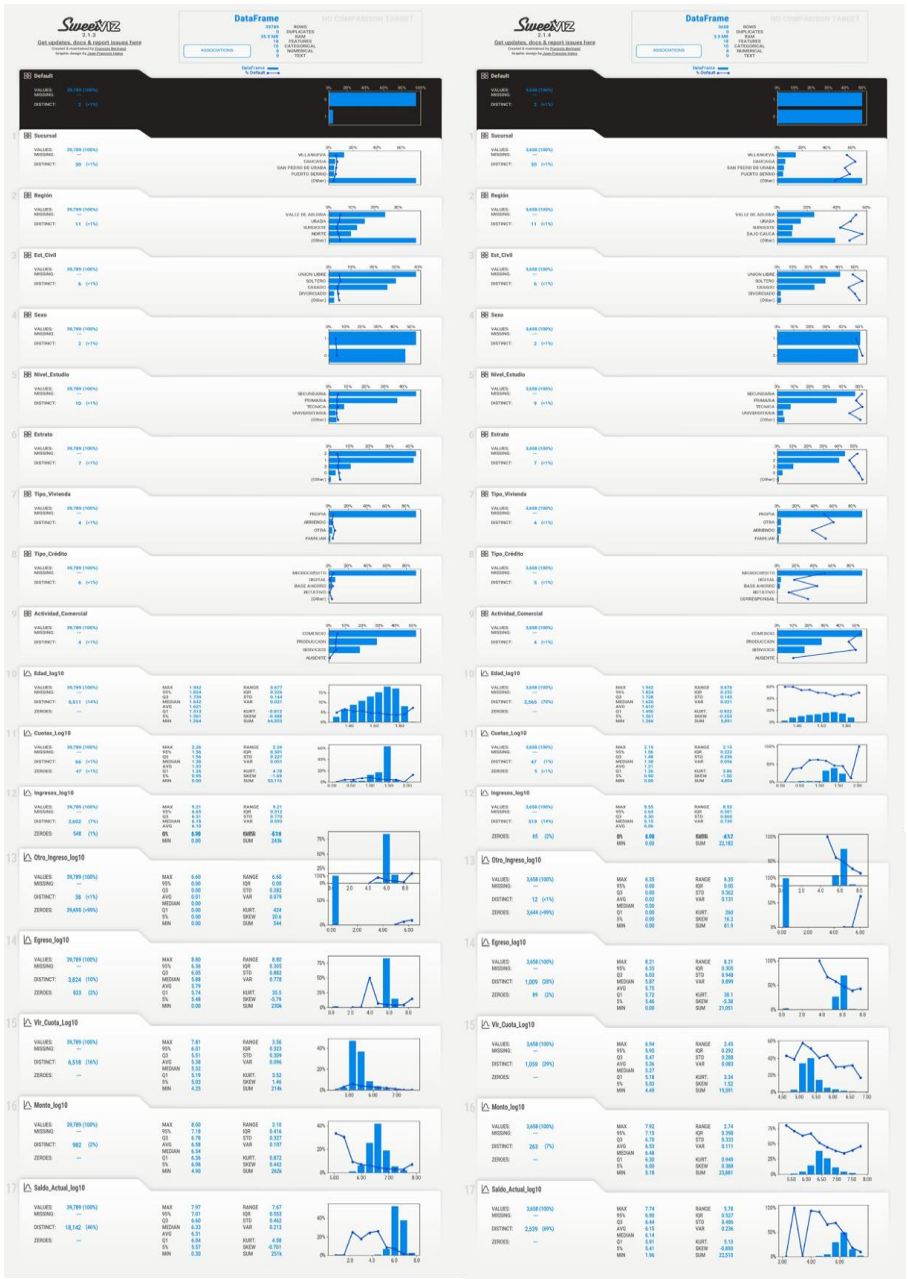
TABLA ORIGINAL

#	Column	Non-Null Count	Dtype
0	Sucursal	39789 non-null	object
1	Región	39789 non-null	object
2	Est_Civil	39789 non-null	object
3	Sexo	39789 non-null	object
4	Nivel_Estudio	37418 non-null	object
5	Edad	39709 non-null	float64
6	Estrato	39789 non-null	int64
7	Tipo_Vivienda	39525 non-null	object
8	Activiad_Ppal	39660 non-null	object
9	Ingresos	39789 non-null	int64
10	Otro_Ingreso	39789 non-null	int64
11	Egreso	39789 non-null	int64
12	Tipo_Crédito	39789 non-null	object
13	Monto	39789 non-null	int64
14	Cuotas	39789 non-null	int64
15	Vlr_Cuota	39789 non-null	int64
16	Última_Cuota	38358 non-null	object
17	Vlr_BDR	39592 non-null	float64
18	Pago_BDR	39789 non-null	int64
19	Se_paga_BDR	39789 non-null	int64
20	Actividad_Comercial	39789 non-null	object
21	Días_Vencido	39789 non-null	int64
22	Saldo_Actual	39789 non-null	int64
	dtypes:	float64(2)	int64(11) object(10)

TABLA IMPUTADA

#	Column	Non-Null Count	Dtype
0	Sucursal	39789 non-null	object
1	Región	39789 non-null	object
2	Est_Civil	39789 non-null	object
3	Sexo	39789 non-null	object
4	Nivel_Estudio	39789 non-null	object
5	Estrato	39789 non-null	object
6	Tipo_Vivienda	39789 non-null	object
7	Tipo_Crédito	39789 non-null	object
8	Actividad_Come	39789 non-null	object
9	Edad_log10	39789 non-null	float64
10	Cuotas_Log10	39789 non-null	float64
11	Ingresos_log10	39789 non-null	float64
12	Otro_Ingreso_lo	39789 non-null	float64
13	Egreso_log10	39789 non-null	float64
14	Vlr_Cuota_Log1(	39789 non-null	float64
15	Monto_log10	39789 non-null	float64
16	Saldo_Actual_lo	39789 non-null	float64
17	Default	39789 non-null	int64
	dtypes:	float64(8)	int64(1) object(9)

Anexo 3. Base de cartera original y balanceado con la técnica del submuestreo.



Anexo 4. Modelos de Clasificación Definidos. Resumen de las métricas promedio reportados por PyCaret

	Model	Accuracy	AUC	Recall	Prec.	F1
MBR	Modelo Basado en Reglas	0,5246	0,5264	0,3089	0,5507	0,3958
dummy	Dummy Classifier	0,9539	0,5000	0,0000	0,0000	0,0000
rf	Random Forest Classifier	0,9428	0,8117	0,0972	0,2231	0,1353
et	Extra Trees Classifier	0,9374	0,7187	0,0758	0,1499	0,1005
lightgbm	Light Gradient Boosting Machine	0,9323	0,8295	0,1546	0,1991	0,1735
dt	Decision Tree Classifier	0,8924	0,5955	0,2685	0,1445	0,1875
gbc	Gradient Boosting Classifier	0,8510	0,8178	0,4562	0,1449	0,2198
knn	K Neighbors Classifier	0,8151	0,6094	0,3122	0,0858	0,1346
ada	Ada Boost Classifier	0,8070	0,7870	0,5320	0,1253	0,2027
lr	Logistic Regression	0,7378	0,7836	0,7237	0,1181	0,2030
lda	Linear Discriminant Analysis	0,7224	0,7722	0,7023	0,1094	0,1893
ridge	Ridge Classifier	0,7223	0,0000	0,7023	0,1094	0,1893
svm	SVM - Linear Kernel	0,6388	0,0000	0,8414	0,1010	0,1793
nb	Naive Bayes	0,1924	0,5773	0,9232	0,0503	0,0954
qda	Quadratic Discriminant Analysis	0,0605	0,5062	0,9971	0,0467	0,0892

Anexo 5. Gráfico representativo de las variables más importantes en el modelo ML. Indexado según su importancia.