Análisis y comparación de modelos de clasificación de aprendizaje automático aplicado a riesgo crediticio

Analysis and comparison of machine learning classification models applied to credit approval

Jorge Brian Alarcón Flores, Jiam Carlos López Malca, Luis Ruiz Saldarriaga, Christian Walter Sarmiento Román

Maestría en Informática con mención en Ciencias de la Computación, Pontificia Universidad Católica del Perú. Lima, Perú.

Recibido el 18 de noviembre del 2017, aceptado el 26 de noviembre del 2017

DOI: https://doi.org/10.33017/RevECIPeru2017.0014/

Resumen

El sector industrial financiero se ha convertido en un sector muy competitivo a nivel mundial. Dentro de este contexto, la decisión del otorgamiento de crédito es uno de los procesos más importantes del cual dependen indicadores críticos del negocio como son las colocaciones, las recuperaciones y el índice de morosidad. Este proceso se ha basado históricamente en expertos del negocio, quienes en base a su experiencia determinaban en función a ciertas variables de comportamiento del solicitante, si debían otorgar o no el crédito. En esta última década, el desarrollo de tecnologías como la inteligencia artificial y el aprendizaje de máquina han aportado mucho en la automatización de este proceso. El presente trabajo tiene como objetivo principal el análisis de varios algoritmos matemáticos basados en el aprendizaje de máquina en las predicciones de otorgamiento de crédito, dando una explicación objetiva de los resultados y sugiriendo las siguientes investigaciones que se desarrollarán con el fin de obtener mejores resultados en los algoritmos matemáticos existentes. Como resultados de la experimentación de determinó que el mejor modelo fue el de Gradient Boosting, con una exactitud de 83.71%.

Descriptores: *inteligencia artificial, aprendizaje automático, riesgo crediticio, modelos matemáticos, gradient boosting.*

Abstract

The financial industry has become into a very competitive sector worldwide. In that sense, the credit granting decision is one of the most important process of all, and in whose accuracy, rests the good performance of several critical business KPI's such as loans level, credit recoveries level and nonperforming loans ratios. This key process has historically based on the experts' judgement, and have taken the decision of granting or not credit loans according to several customer credit behavior elements. In the last decade, the developing of certain technology such AI and machine learning has allowed this process automation. The present paper has its main goal, the analysis of several mathematical algorithms based on machine learning and the exposition of which of them have the better results in credit granting predictions to collaborate with current knowledge in this particular issue, giving an objective explanation of the results and suggesting following researches to be developed in order to get better results in existing mathematical algorithms. As results of the experimentation determined that the best model was Gradient Boosting, with an accuracy of 83.71%.

Keywords: *artificial intelligence, machine learning, credit risk, mathematic models, gradient boosting.*

1. Introducción

Las instituciones crediticias deben establecer esquemas eficientes de administración y control del riesgo de crédito al que se exponen en el desarrollo del negocio, en resonancia a su propio perfil de riesgo, segmentación de mercado, según las características de los mercados en los que opera y de los productos que ofrece; por lo tanto, es necesario que cada entidad desarrolle su propio esquema de trabajo, que asegure la calidad de sus portafolios y además permita identificar, medir, mitigar y monitorear las exposiciones de riesgo de contraparte y las pérdidas esperadas, a fin de mantener una adecuada cobertura de provisiones o de patrimonio técnico.

Para ello, es necesario que se adopte un procedimiento de investigación y análisis, el cual se logre ver reflejado en un scoring de crédito o evaluación del riesgo crediticio, considerándose como uno de los procesos más importantes de toda institución financiera, que pueda permitir tener un primer acercamiento del cliente en el proceso de admisión del crédito, el cual pueda otorgar a la institución financiera, un elemento de ventaja competitiva dentro del sector industrial, pues a partir de éste, es posible mejorar indicadores clave del rendimiento del negocio, tales como las colocaciones, las recuperaciones y el índice de morosidad. El problema busca por lo tanto trabajar un modelo matemático basado en aprendizaje automático que permita predecir de manera acertada y con un alto porcentaje de exactitud la calificación de un cliente (Bueno o Malo), el cual nos brinde un elemento adicional para la toma de decisión, el cual será finalmente, el otorgamiento o no otorgamiento del crédito.

2. Objetivo del estudio

El objetivo principal de esta investigación es encontrar el mejor modelo de clasificación que permita encontrar una regla óptima de decisión de solicitud de créditos al Banco Alemán, en base a las características de los solicitantes y las solicitudes que realizan.

La buena elección de los clientes a quienes se les otorgará un crédito significará una reducción de costos y menores riesgos para el banco. Se aplicarán diferentes modelos de clasificación al conjunto de datos con la finalidad de comparar los resultados y determinar cuál presenta una mejor predicción. La utilización de diversos modelos de clasificación, aparte de buscar el que más se ajuste al problema descrito, permitirá cumplir con el objetivo general del curso, que es la aplicación en el diseño, implementación y evaluación de algoritmos de aprendizaje de máquina. El contacto con diversos modelos entrega la posibilidad de aprender las características de cada uno de ellos, así como conocer las ventajas y desventajas de los mismos.

3. Estado del arte

En relación con estudios que han utilizado el conjunto de datos seleccionado en esta investigación, se mencionan los siguientes:

Combining Feature Selection and Neural Networks for Solving Classification Problems (2001). Este estudio fue realizado por el Departamento de Tecnologías de la Información de la Universidad Nacional de Irlanda. Inicialmente, desarrolla la aplicación del modelo de redes neuronales y considera el total de características otorgadas en el conjunto de entrenamiento (20 características) y divide el total de muestras (1000) en dos conjuntos: Entrenamiento (666 muestras) y test (334 muestras). Luego se intenta mejorar esta predicción aplicando el algoritmo de selección de características, donde se reduce el número original de atributos a 7 y nuevamente se aplica el modelo para comparar los resultados. El resultado mostró que la exactitud de los datos de entrenamiento disminuyó cuando se eliminaron 13 atributos de los datos de entrada. Sin embargo, la precisión predictiva fue marginalmente mayor con la exclusión de los atributos redundantes.

Application of Artificial Intelligence (Artificial Neural Network) to Assess Credit Risk: A Predictive Model for Credit Card Scoring (2009). Este estudio fue desarrollado en el Blekinge Instituto de Tecnología de Suecia. En esta investigación se trabajó el conjunto de datos, con la aplicación de los modelos de análisis discriminante, regresión logística y redes neuronales. El modelo con el que se obtuvo mejores porcentajes de aciertos de clasificación es el de redes neuronales, con un 83.86%, mientras que con el porcentaje de aciertos de clasificación con los modelos de análisis discriminante y regresión logística fue de 76.40%.

Respecto a investigaciones que han abordado el mismo problema en estudio, pero con otros conjuntos de datos, se mencionan los siguientes:

Bank credit risk analysis with k-nearest neighbor classifier: Case of Tunisian bank. (2015). En esta investigación se estudió el problema del riesgo crediticio aplicado a una base de datos de 924

registros de crédito concedido por un banco comercial tunecino entre los años 2003 y 2006. El modelo de clasificación de aprendizaje automático utilizado para estudiar este problema fue el de K-Nearest Neighbors (K-NN). Se probó el modelo utilizando diferentes valores de K (2, 3, 4 y 5). El criterio utilizado para evaluar el rendimiento es la minimización de la tasa de riesgo. Los principales resultados muestran que el mejor K-NN obtenido es con k = 3, con un porcentaje de clasificación global de 88.63%. Además, para evaluar el rendimiento de la curva modelo se representa ROC. El resultado muestra que el criterio AUC (Area Under Curve) es del orden del 95,6%.

4. Diseño del experimento

4.1. Descripción del conjunto de datos

El conjunto de datos original consta de 1,000 observaciones y 20 atributos, además del clasificador. Las observaciones y características se resumen de la siguiente manera:

- Número de observaciones: 1000.
- Número de muestras: Para el entrenamiento 750 observaciones y para la prueba 250 observaciones.

Las características contienen valores discretos y continuos, y dentro de los discretos, existen variables numéricas y categóricas. La clasificación (target) tiene dos valores; a saber, 1 si el crédito es aprobado o 2 si el crédito se rechaza. El conjunto de datos está claramente desbalanceado, pues existen 700 instancias con valor de clasificación 1 (crédito aceptado) y 300 con clasificación 2 (crédito rechazado).

4.2. Estrategia de preparación de los datos

Uno de los primeros problemas que se va a enfrentar es preparar la data para el entrenamiento y validación de los distintos modelos de aprendizaje a utilizar. Se deben resolver los siguientes puntos:

4.2.1. Estrategia de balanceo del conjunto de datos

En una primera instancia se quiso balancear el conjunto de datos con la técnica de oversampling. Esto significaba duplicar aleatoriamente 400 instancias correspondiente al valor 2 del target (rechazo del crédito) de tal manera de llegar a 700 instancias con aprobación de crédito y 700 instancias con rechazo del crédito. Sin embargo, la teoría existente no la considera recomendable debido a que el oversampling excede al total de instancias totales de rechazo de crédito (valor 2). Esto hubiese significado que un mismo punto (x,y) pudiese existir hasta 3 veces (uno original y hasta dos copias producto del oversampling). La otra alternativa hubiese sido hacer un undersampling de los valores del target de aprobación del crédito, llevándolo de 700 a 300 instancias, para igualar las instancias de ambas clasificaciones (300 positivas y 300 negativas). Esta estrategia hubiese significado descartar más del 50% de datos de la clasificación de aprobación del crédito, lo cual tampoco consideramos adecuado. Por tal motivo se optó por una estrategia mixta (oversampling y undersampling) a fin de "suavizar" el impacto del balanceo. Se aplicó por lo tanto undersampling de la clasificación positiva llevando de 700 a 500 instancias y oversampling de la clasificación negativa llevándola de 300 a 500 instancias. De esta manera se mantuvo un conjunto de datos de 1,000 instancias con 500 de cada clasificación.

Luego de las primeras experimentaciones y en función a la teoría existente, se encontró que existe una técnica de oversampling más eficiente. Esta técnica es SMOTE. La ventaja de esta técnica es que permite generar nuevas observaciones (x,y) en función a la distribución de las características del conjunto de datos. En función a esta técnica, se mantuvieron las 700 instancias positivas provenientes del conjunto de datos originales y generar 300 observaciones adicionales de la clase minoritaria (rechazo del crédito) hasta completar 700. De esta manera el conjunto de datos quedó balanceado con 1,400 instancias (700 de cada clase).

Por lo tanto, la estrategia inicial para el balanceo de los datos fue replanteada con la técnica de SMOTE. Al final se experimentó con el conjunto de datos modificado de 1,400 instancias balanceado a 700 instancias por cada clase.

4.2.2. Análisis y categorización de las características existentes

Con respecto a las variables numéricas, se ha trabajado la conversión de sólo dos variables, estas son: Monto del crédito (Credit amount), donde se procedió con la división por 1000 y Edad (Age), para este variable se definieron rangos, estas conversiones permitirán reducir la amplitud de dichas variables, con respecto a las demás variables numéricas se decidió mantener sus valores originales.

Para el caso de las variables categóricas se aplicó la técnica del one-hot encoding para transformar los

datos que no son ordinales en datos numéricos binarios. Esto permite la multiplicación de las características, pasado de 20 del conjunto de datos original a 62 bajo esta técnica.

4.2.3. Algoritmos utilizados

Los algoritmos o modelos seleccionados para las pruebas son los siguientes:

- Regresión Logística
- Redes Neuronales
- Support Vector Machine
- Random Forest
- Gradient Boosting

4.2.4. Selección y justificación de la medida de calidad

Se seleccionó dos medidas de calidad con el objetivo de poder comparar la validez y eficacia de los distintos modelos que se van a utilizar. El primer indicador de calidad es la exactitud de cada modelo. Se eligió este índice; a pesar de que no es conveniente utilizarlo en conjunto de datos desbalanceados, porque basados en la estrategia de balanceo que hemos planteado queremos tenerlo como referencia y comparación entre los distintos modelos. El segundo indicador de calidad elegido es la curva ROC, el cual presenta buen funcionamiento analizando el comportamiento de clasificadores en todos los umbrales posibles, lo cual nos permitirá medir cuán óptimos son cada uno de los modelos desarrollados bajo los escenarios del conjunto de datos en su estado original (desbalanceo) y con su posterior tratamiento aplicándosele técnicas de balanceo.

4.2.5. Selección de las características más relevantes para el aprendizaje del modelo

Una vez seleccionado el mejor modelo de clasificación, se realizó el análisis de importancia de variables, con el fin de encontrar aquellas características de mayor contribución en la predicción del Riesgo crediticio.

El modelo seleccionado para esta investigación es el Gradient Boosting, el cual presenta una medida de importancia de variables basada en el promedio del número de veces que se ha seleccionado cada variable en cada partición, ponderado por la mejora al cuadrado del modelo como resultado de cada partición. [14]

$$\widehat{I}_j^2(T) = \sum_{t=1}^{J-1} \widehat{\imath}_t^2 \, 1(v_t = j)$$

Donde la sumatoria es sobre los nodos no terminales t del nodo terminal J del árbol T. v_t es la variable de division asociada con el nodo t y $\widehat{\imath}_t^2$ es la correspondiente mejora empírica en error cuadrático. [15]

5. Discusión y resultados

Como se puede observar en la Tabla 1, en la experimentación, el modelo con mayor exactitud fue el de SVM con Kernel Gaussiano. Sin embargo, hemos comprobado que no es recomendable usar este algoritmo cuando se cuenta con un número limitado de características en comparación con el número de observaciones del conjunto de datos de entrenamiento. El motivo de esta conclusión es debido a que pensamos que existe un riesgo de overfitting del modelo, debido al bajo ratio observaciones/características, teniendo en cuenta que una baja cantidad de observaciones en comparación a la cantidad de características no permite un ajuste adecuado del modelo frente a la dispersión de puntos causado por la cantidad de características. Por tal motivo no lo estamos considerando.

Tabla 1: Comparación de algoritmos utilizados

Modelo	Datos Originales	Datos balanceados (SMOTE)	Datos Encoding	Datos Encoding Tuned
SVM - Kernel	0.732	0.763	0.794	0.860
Gradient Boosting	0.776	0.820	0.820	0.837
Random Forest	0.776	0.811	0.817	0.809
Redes Neuronales	0.760	0.766	0.780	-
Regresión Logística	0.748	0.746	0.746	-

*Los modelos de redes neuronales y de regresión logística no fueron considerados en el escenario de Datos Encoding Tuned, debido a la menor precisión obtenida en los primeros escenarios desarrollados.

El modelo de clasificación Gradient Boosting es el que mejor regularidad tuvo durante las experimentaciones. En la Figura 1 observamos la curva de aprendizaje en el mejor de los escenarios,

obteniendo su menor error con numero de nodos base = 200.

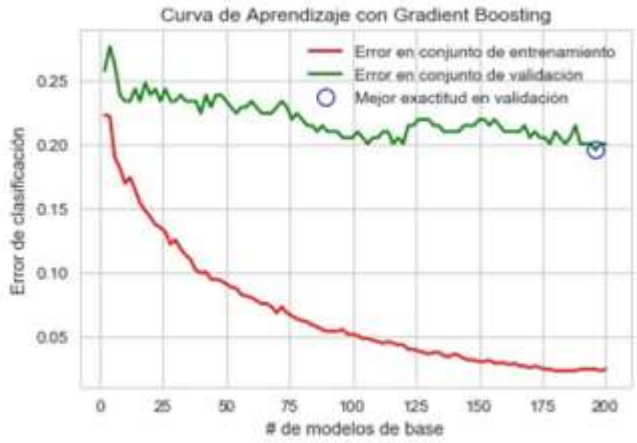

Figura 1. Curva de Aprendizaje con Gradient Boosting

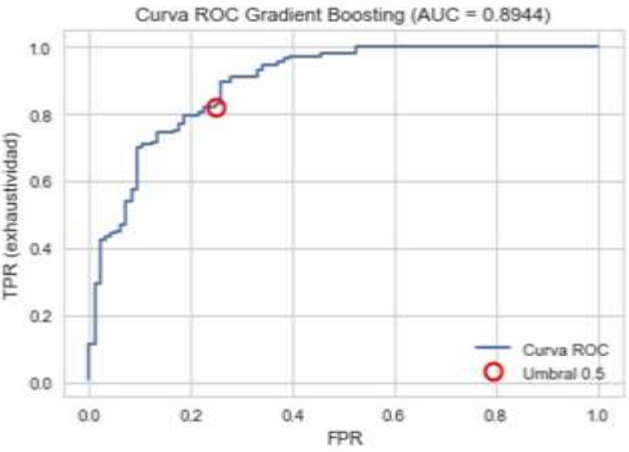

Figura 2. Curva ROC Gradient Boosting

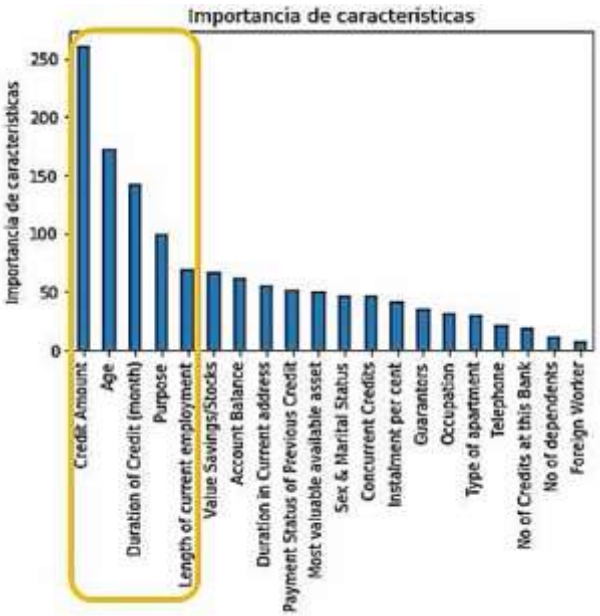

Figura 3 Importancia de variables Gradient Boosting

6. Conclusiones y trabajos futuros

Para el proceso de experimentación se concluye que el tratamiento de los datos previamente al uso de los modelos es de suma importancia para un mejor rendimiento del modelo. En este caso, el balanceo y el análisis de cómo transformar las características aportó en las mejoras incrementales del rendimiento. Las variables numéricas aportan más al modelo cuando se acotan (reducir dispersión), las variables categóricas también mejoran el modelo cuando se convierten a numéricas binarias analizando además si el orden de los valores discretos influye o no en el modelo.

Uno de los parámetros de calidad para medir los modelos es la exactitud. Desde el punto de vista del negocio una mala clasificación (falsos positivos o falsos negativos) genera costos asociados. En este caso del otorgamiento de crédito, un falso positivo significa otorgarle crédito a alguien que no está calificado, pudiendo generar problemas de cobranza, y en el caso de los falsos negativos implica negarle un crédito a alguien que sí está calificado teniendo implicancias con el nivel de colocaciones y los ingresos de la institución financiera.

En la experimentación, el modelo con mayor exactitud fue el de SVM con Kernel Gaussiano. Sin embargo, hemos comprobado que no es recomendable usar este algoritmo cuando se cuenta con un número limitado de observaciones en comparación con el número de características del conjunto de datos de entrenamiento. Se comprueba además que, en los distintos valores de la experimentación de este modelo, existe un incremento importante en la exactitud del último escenario, lo cual podría deberse a un overfitting del modelo. Por tal motivo no lo estamos considerando. Dado lo anterior, el modelo que se comportó de manera más estable con los distintos escenarios de experimentación fue el de Gradient Boosting con una exactitud de 83.71%.

Las características de mayor importancia para el modelo Gradient Boosting fueron el saldo de la cuenta (balance account), duración del crédito (duration of credit), monto del crédito (credit amount), antigüedad laboral (length of current employment) y la edad (age). Todas estas variables son coincidentemente las que los expertos del sector financiero toman en consideración al momento de evaluar el otrogamiento de los créditos. Esto muestra una concordancia entre el mundo real y los modelos matemáticos.

Es una buena práctica la aplicación de la técnica de grid search para la mejora del modelo. En el presente trabajo de investigación, se obtuvo una mejora en la exactitud de los modelos aplicando grid search buscando los parámetros óptimos para cada modelo.

Como investigaciones futuras al presente trabajo se sugiere en primer lugar comprobar el motivo por el cual el modelo Kernel Gaussiano presentó un salto importante en su exactitud en el último escenario de experimentación que consistió en el uso de la técnica del grid search luego de aplicar one-hot encoding de las variables categóricas. La explicación a priori del presente trabajo es que el modelo en este escenario está generando overfitting.

Existen técnicas que se podrán aplicar para continuar mejorando la exactitud, por ejemplo, se puede realizar un análisis de características importantes o reducir las características del modelo usando forward o backward elimination. Se cree que para este caso de estudio backward elimination es la mejor opción, pero esto se dejará para una experimentación futura. Por último, es también posible complementar la presente experimentación usando stacking de varios modelos en conjunto con la finalidad de mejorar los resultados de la exactitud.

Referencias

[1] N. Ghatasheh, *Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study*, International Journal of Advanced Science and Technology. **72** (2014) 19-30.

[2] C. L. Huang, M. C. Chen y C. J. Wang, *Credit scoring with a data mining approach based on support vector machines*, Expert Systems with Applications. **33** (2007) 847-856.

[3] J. A. Bastos, *Credit scoring with boosted decision trees*, MPRA Paper. **27** (2008). 262-273.

[4] S. Dahiya, S.S Handa y N.P. Singh, *Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set*, Industrija. **43** (2015) 163-174.

[5] C. Josphat Kipchumba, *Credit evaluation model using Naive Bayes classifier: A Case of a Kenyan Commercial Bank*, MSc thesis, University of Nairobi, 2012.

[6] H. Van Sang, N. Ha Nam y N. Duc Nhan, *A novel credit scoring prediction model based on Feature Selection approach and parallel random forest*, Indian Journal of Science and Technology. **9** (2016).

[7] B. Baesens, R. Setiono, C. Mues and J. Vanthienen, *Using Neural Network rule extraction and decision tables for credit risk evaluation*, Computer Journal of Management Science. **49** (2003) 312-329.

[8] C. Lakshmi Devasena, *Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction*, IJCA Proceedings on International Conference on Communication, Computing and Information Technology ICCCMIT. **3** (2014) 30-36.

[9] M. Haltuf, *Support Vector Machines for Credit Scoring* (Vysoká škola ekonomická v Praze, 2013).

[10] P. O'Dea, J. Griffith, C. O'Riordan, *Combining feature selection and Neural Networks for solving classification problems (Technology Department*, National University of Ireland, 2001).

[11] S. Islam, L. Zhou y F. Li, *Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring* (MSc thesis, School of Management, Blekinge Institute of Technology, 2009).

[12] A. K. Abdelmoula, *Bank credit risk analysis with knearest Neighbor Classifier: Case of Tunisian Banks*, Accounting and Management Information Systems. **14** (2015) 79-106.

[13] M. Kern and B. Rudolph, *Comparative Analysis of Alternative Credit Risk Models–an Application on German Middle Market Loan Portfolios* (CFS Working Paper Series, Universität Frankfurt a. M., 2001).

[14] J. Elith, J. R. Leathwick y T. Hastie, *A working guide to boosted regression trees*, Journal of Animal Ecology. 77 (2008) 802-813.

[15] J. H. Friedman, *Greedy function approximation: A gradient boosting machine*, The Annals of Statistics, **29** (2011) 1189-1232.

E-mail: brian.alarcon@pucp.edu.pe; jiam.lopez@pucp.edu.pe; l.ruiz@pucp.pe; cwsarmiento@pucp.edu.pe