Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Grandes volúmenes de datos y riesgo de crédito: técnicas de machine learning para el default en tarjetas de crédito

AUTOR: STEFANO NICOLAS BELTRAME

2020

.**Agradecimientos**

Si bien este trabajo podría llegar a resultar “simple”, finalizarlo con éxito y atrasándome tan solo un par de meses con respecto a la fecha en la cual debería haber sido presentado, no fue cosa fácil. Desde agosto 2019 tuve la suerte poder comenzar un nuevo camino profesional en Barcelona (Cataluña, España), que seguramente muchos no sabrán, es mi ciudad natal. El anhelo de volver a la tierra donde nací y en parte me crie fue tan grande, que con ayuda del destino y la maravillosa coordinación de diferentes cuestiones de carácter “azarosas” (aunque a esta altura no considero que así lo sean), es que hoy me encuentro escribiendo estas palabras de agradecimiento a un conjunto de seres queridos.

En primer lugar, me gustaría agradecer a mis padres, **Dra. Maria Cecilia Tenaglia** e **Ing. Renato Beltrame**, ya que fueron los que me apoyaron económica, mental y afectivamente a seguir estudiando una vez terminada la carrera de grado este tipo de posgrado.

También, quisiera agradecer a mi psicóloga, **Lic. Florencia Rosetti**, ya que sin ella no hubiera sido capaz de lograr un correcto balance entre estudio, trabajo y vida social a lo largo de toda la Especialización, tanto para la cursada presencial en Argentina como para la continuación de mis estudios en España.

Al **Mg Joaquín Bossano**, un gran amigo al cual le agradezco todos los comentarios que realizó sobre este trabajo dado su expertise en esta temática.

Al **Director Académico** de esta especialización **Dr. Javier García Fronti**, que me permitió, sin ningún problema, a continuar mis estudios a la distancia desde Barcelona

Al **Coordinador Académico** de esta especialización Dr. Pablo Herrera, mi tutor de Trabajo Final, que, sin su apoyo y presencia desde mi arribo a Barcelona, este trabajo jamás se hubiese finalizado y hoy no habría accedido a este título de especialista.

Al Profesor de esta especialización **Mg Rodrigo del Rosso**, por sus excelentes y clarísimas clases de programación de R. Sin su compromiso para con la docencia, este trabajo jamás hubiese sido programado en dicho código de programación.

Al Profesor de esta especialización **Mg. Roberto Abalde** que, sin su particular pasión por la docencia y su forma de ser tan directa, jamás me hubiera animado a decir que “si” a mi trabajo actual, que con mucha pasión y cariño realizo todos los días.

Al **Mg. Juan David Ossa Gómez** que sin haberme contactado por primera vez en LinkedIn posiblemente no hubiera elegido este tema tan interesante ni quizás hubiera decidido llevar adelante este posgrado.

Al **Mg. Matías Lencina**, que con toda su sabiduría sin saber nada del tema, me hizo ver el camino de la luz con sus infinitos consejos para la vida, incluidos aquellos relacionados a la finalización de estos estudios.

A mis **compañeros de la Especialización** en general, pero en particular al **Actuario Pablo Silvarredonda**, que nunca dejo de apoyarme para que no deje mis estudios desde que llegué a Barcelona. No tengo palabras de agradecimiento de todas las cosas en las cuales me ayudo desde que llegue aquí: apuntes, libros, horas por Skype explicándome las cosas que no entendía, y obviamente apoyándome emocionalmente cada vez que sentía que no podía avanzar para finalizar este posgrado.

Por último, no puedo dejar de mencionar a mi querido amigo **Nicolas Mackinlay, Lic. en Economía** y compañero de estudio durante gran parte de la carrera de grado. De no habérmelo cruzado sentado en ese banco de la facultad antes de entrar a cursar, jamás estaría escribiendo estas palabras de agradecimiento hacia tanta gente. Él, es el gran responsable de mi salto personal y profesional. Gracias a él pude ponerme en contacto con gente en Barcelona, responsables del trabajo que tengo actualmente, antes de saber que existiese una remota chance de poder venirme a vivir hacia aquí.

Resumen

En el presente trabajo se utilizaron diferentes técnicas de minería de datos para realizar una clasificación binaria de 30.000 clientes en tarjetas de crédito taiwanesas, entre los meses abril y septiembre del año 2005, para determinar si defaultearán o no sus pagos con tarjeta al mes siguiente, dada una cierta confianza que eso sea cierto.

Para ello, se entrenaron cinco algoritmos de clasificación (Decision Tree, K-NN, Random Forest, Redes Neuronales y Regresión Logística) utilizando como métrica para determinar la calidad de los modelos el área bajo la curva ROC. Los resultados de esta investigación sugieren que, para la base datos tomada como ejemplo, el algoritmo Random Forest es aquel que maximiza el área bajo la curva ROC, siendo dicho resultado muy superior al baseline.

Palabras claves: Minería de datos, riesgo crediticio, score de crédito, default, tarjetas de crédito taiwanesas, área bajo la curva ROC.

Estructura

Resumen 5
Introducción 7
Definiciones, marco conceptual y literatura relacionada 10
1.1. Grandes volúmenes de datos: Big Data y la gestión del riesgo crediticio 10
1.2. Marco teórico y conceptual 11
1.3. Literatura Relacionada 13
Revisión y análisis exploratorio del data set utilizado 15
2.1. Fuentes de datos 15
2.2. Análisis exploratorio sobre las variables continuas 18
2.3. Análisis exploratorios sobre las variables discretas 22
Modelos clasificatorios para estimar el grado de default 25
3.1. Algoritmos de minería de datos y métricas de evaluación utilizadas 25
3.2. Aplicación del modelo a través de técnicas de aprendizaje automático 26
3.3. Resultados obtenidos e interpretación de resultados 27
Conclusiones y Futuros Estudios 29
4.1. Conclusiones 29
4.2. Futuros estudios 30
Referencias bibliográficas 31

Introducción

Una de las definiciones de *Big Data* viene dada por "*conjunto de grandes volúmenes de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados*" (Laney, 2001). Así, el término hace referencia a la necesidad de desarrollar nuevas tecnologías y metodologías para poder extraer información valiosa a partir de dichos volúmenes de datos, con el propósito de tomar mejores decisiones por parte las organizaciones.

Entre estas nuevas metodologías se encuentra la minería de datos, que sirve para extraer información estratégica "escondida" en grandes bases datos. Entre sus aplicaciones se encuentra el riesgo de crédito, ya que provee información para comprender de mejor manera el mercado crediticio y ayuda a los analistas a tomar mejores decisiones (ej: a quien prestar y a quien no).

Un ejemplo de esto es la clasificación binaria de clientes de acuerdo con su probabilidad de default (segmentación avanzada), siempre y cuando se cuente con la información socioeconómica para hacer esa clasificación. Esto segmenta a los clientes en dos categorías ("buenos" con baja probabilidad de default y "malos" con alta probabilidad de default) lo cual incrementa la eficiencia con la que se asigna el crédito. En esta misma línea, la clasificación de clientes puede ser aplicada para detectar y evitar el fraude en las aplicaciones de crédito entre clientes de una institución financiera, lo cual ahorra pérdidas potenciales para cualquier organización de este tipo.

En un contexto donde abundan los problemas de información asimétrica, y donde está en juego una inmensurable masa de riqueza, este trabajo tiene por objetivo desarrollar un modelo de minería de datos que permita la correcta clasificación en dos categorías (default y no default) de los clientes de acuerdo con su probabilidad de impago, utilizando como ejemplo una base de datos de tarjetas de crédito taiwanesa. En otras palabras, esta investigación permitiría colaborar a mejorar los perfiles de solvencia crediticia para los clientes de una institución financiera, dando conjeturalmente como resultado una reducción de las perdidas esperadas asociadas al riesgo de crédito.

Vale la pena mencionar que este trabajo utiliza como *input* una base de datos de carácter pública, perteneciente a "UCI Machine Learning Repository". Teniendo en cuenta que este tipo de información suele ser confidencial, se decidió elegir un data set de libre

acceso con el fin de evitar el quebranto de cualquier ley contenida en el código penal. En otras palabras, la elección de dicho data set es en parte *trivial*, ya que el foco del trabajo está puesto en los métodos clasificatorios aplicados al riesgo de crédito en general y no al data set en particular – no se considera relevante ni el tipo de producto crediticio ni donde se generaron los datos.

En consecuencia, la pregunta de investigación que conduce el presente trabajo es: **¿Cuáles técnicas de minería de datos permiten realizar una mejor clasificación binaria de clientes en tarjetas de crédito para determinar si defaultearán o no sus pagos al mes siguiente, dada una cierta confianza que eso sea cierto?**

Con el fin de realizar una conclusión integral a la pregunta formulada, el presente trabajo tiene como **objetivo general** desarrollar un modelo de minería de datos que permita la correcta clasificación en dos categorías (default y no default) de los clientes de acuerdo con su probabilidad de impago, utilizando como ejemplo una base de datos de tarjetas de crédito taiwanesas. Para poder lograr este objetivo general, se plantean los siguientes **objetivos específicos**:

1. Establecer el marco teórico/conceptual de dicho trabajo.
2. Revisar la literatura relacionada con el riesgo de crédito.
3. Realizar el correcto análisis exploratorio de las variables contenidas en la base de datos.
4. Desarrollar diferente tipo de modelos clasificatorios.
5. Analizar las características del modelo que clasifica mejor a los clientes de acuerdo con su probabilidad de default.

Para lograr un mejor entendimiento del análisis realizado, en un **primer capítulo**, se hace foco en marco teórico sobre que se considera riesgo crediticio. Asimismo, se estudia el impacto de los grandes volúmenes de datos sobre las instituciones otorgadoras de crédito. Se profundiza respecto del cambio de paradigma que enfrentan las instituciones crediticias ante la presencia disruptiva de los "Big Data". Por último**,** se presenta un breve resumen de la literatura relacionada con los modelos de riesgo crediticio (o de probabilidad de default).

Luego, en un **segundo capítulo,** se realiza el análisis descriptivo y exploratorio de las variables contenidas en el data set, que son utilizados para calibrar correctamente los algoritmos entrenados en el **tercer capítulo**. En este capítulo se desarrolla sobre los diferentes modelos de

clasificación de clientes, haciendo foco en las técnicas de minería de datos, las métricas de evaluación, y la interpretación de los algoritmos de clasificación para entrenar los diferentes modelos.

Por último, pero no menos importante, el **cuarto capítulo** contiene la **conclusión** de todo el trabajo, resumiendo los principales resultados obtenidos como las conclusiones que se desprenden de dicho. También se presentan las **recomendaciones para futuros** estudios que pueden ser realizados con esta misma base.

Definiciones, marco conceptual y literatura relacionada

Los principales proyectos en donde las entidades bancarias han aplicado estas nuevas técnicas fueron a la gestión la prevención de fraude y de riesgos crediticios. Por ello, resulta pertinente hacer una revisión del impacto que tienen las nuevas tecnologías en el manejo y gestión del riesgo crediticio en una institución financiera. En este capítulo se analizarán los principales hitos de los grandes volúmenes de datos sobre el riesgo creditico. Asimismo, se revisa brevemente que se considera riesgo de crédito y se revisa la literatura relacionada con los modelos que se encargan de estimar el riesgo crediticio.

1.1. Grandes volúmenes de datos: Big Data y la gestión del riesgo crediticio

Si bien existen diferentes definiciones sobre lo que es – y lo que no es – Big Data, en este trabajo adoptamos el enfoque de Laney (2001) donde los considera un "*conjunto de grandes volúmenes de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados*". En este artículo, el autor define a los Big Data con relación a sus características principales lo cual se conoce como *"el desafío de las 3 V"*: *Volumen, Velocidad* y *Variedad*. Si bien hace un par de años Laney (2011) incorpora también a las características 2 "V" adicionales – *Veracidad* y *Valor*, en este trabajo solamente nos concentraremos en describir las primeras 3.

La primera fue definida previamente y es aquella en donde los softwares habituales como Excel o el lenguaje SQL no alcanzan para procesarlos, por lo que se requieren nuevos sistemas para poder tratarlos correctamente (ejemplo: NoSQL, Hadoop, entre otros). La segunda hace referencia a que el proceso generador de los datos es *real time*.

Por último, la tercera hace referencia a las diversas formas que pueden tomar los datos: estructurados y no estructurados. Los primeros son aquellos que pueden ser almacenados en tablas relacionales, con longitudes y formatos bien definidos. Los segundos carecen de formatos específicos y deben ser tratados de una mera específica para poder ser almacenados correctamente, siendo en este último caso donde entra en juego la definición de *Big Data* previamente establecida.

Actualmente, una de sus aplicaciones es a la gestión de riesgos crediticios. En particular, la correcta utilización de los grandes volúmenes de datos permite establecer modelos de predicción del riesgo más potentes. Asimismo, los datos pueden ser analizados casi en tiempo real por lo que el tiempo de respuesta es mucho más rápido y el de reacción disminuido. Además, los Big Data permiten la automatización de proceso, mayor precisión de los sistemas predictivos, menor riesgo de error, etc. (Redacción Byte TI, 2017).

Entre las principales técnicas para tratar los datos no estructurados, se encuentra la utilización de la minería de texto – o "*text mining*" que surge como una técnica capaz de convertir el texto no estructurado en datos estructurados, extraer índices numéricos significativos del texto y, por lo tanto, hacer que la información contenida en el texto sea accesible a los diversos algoritmos de minería de datos (Munafo, 2019). Así, muchas entidades financieras utilizan datos no estructurados provenientes de redes sociales como enfoque alternativo de puntuación de crédito y mejorar su evaluación del riesgo crediticio (Munafo, 2019).

El hecho de poder convertir el sentimiento en texto permite mejorar los modelos calificación crediticia de los bancos y mejorar problemas estructurales que tienen los modelos con los cuales trabajan (por ejemplo, que el *score* o probabilidad de default se calcula en base a datos que tiene un rezago en función de la nueva información generada).

Por su parte, las normativas vigentes que deben cumplir las instituciones financieras con relación a los parámetros a modelizar para estimar el riesgo crediticio (Acuerdos de Basilea, II y III) no incluyen este tipo de herramental tecnológico, por lo que se hace difícil saber el potencial impacto que pueden tener. En este mundo entran las nuevas *Fintech* – aquellas entidades financieras que determinan el score de crédito y la tasa a la cual prestan en función de modelos de aprendizaje automático, muchas veces que incluyen este tipo de técnicas.

Será cuestión del paso del tiempo para poder ver cuál es la decisión que tome el BIS – *Bank of International Settlements* sobre él marco de regulación que acogerá a este nuevo tipo de entidades financiera cuyo crecimiento en cantidad y tipo de servicios que ofrecen ha venido creciendo a lo largo del tiempo.

1.2. Marco teórico y conceptual

Conceptualmente, existen diferentes tipos de riesgos que, de no ser correctamente administrados, pueden generar pérdidas económicas de gran magnitud para este tipo de entidades. Entre ellas se encuentran: el riesgo de fraude; de blanqueamiento de capitales; de crédito comercial, operacional, integrado y crediticio. Este trabajo solamente se concentrará en el último debido a que es aquel relevante para aproximar a su objetivo general.

El riesgo crediticio puede ser definido como "*la posibilidad de pérdida debido al incumplimiento del prestatario o la contraparte en operaciones directas, indirectas o de derivados que conlleva el no pago, el pago parcial o la falta de oportunidad en el pago de las obligaciones pactadas"* (Superintendencia de Bancos y Seguros República del Ecuador, 2003).

Así, a lo largo de la historia, el crédito se ha convertido en el instrumento donde los bancos y/o entidades financieras fueron capaces de hacer crecer sus ingresos a niveles muy elevados. Sin embargo, la acción de prestar y tomar prestado puede traer aparejado un incremento en las pérdidas esperadas para dichas instituciones siempre y cuando el riesgo crediticio en general, y en particular el asociado al de la solvencia de sus clientes, no sea correctamente administrado. Por ejemplo, muchas veces las instituciones financieras al no poder distinguir entre los diferentes tipos de clientes (entre los "buenos" pagadores y "malos" defaulteadores) se terminan inclinando hacia aquellos que les son adversos a sus propios intereses ("selección adversa", Akerlof 1970).

Por este motivo, si los bancos tienen indicios de que un préstamo no sea repagado (debido a que hay retrasos en sus pagos), la pérdida debería ser reconocida y ser enviada a *previsión* (o también llamadas reservas). Por ello, los bancos calculan la *perdida esperada* de un préstamo de acuerdo con el enfoque contenido en el Segundo Acuerdo de Basilea para intentar dar cuenta del potencial riesgo crediticio que puedan tener en sus carteras. En él, la pérdida esperada surge del producto entre la *probabilidad de default* (la probabilidad de que un préstamo no sea devuelto), *el monto de la perdida al momento del default* (es decir, cuanto del préstamo se perdió al momento del default), y la *exposición al default* (es decir, el valor del préstamo al momento del default.).

En este marco, dicho trabajo considera *solamente* la probabilidad de default para determinar el riesgo crediticio o riesgo de impago de un agente económico. Asimismo, si bien empíricamente cambios en el *score* crediticio – calificación – de algún instrumento,

persona física y/o jurídica podrían no implicar cambios en la probabilidad de default, en este trabajo se asumen que correlacionan uno a uno.

Por su parte, el riesgo crediticio varía en función del tipo de producto que se comercialice. En particular, uno de los mayores riesgos de crédito emerge del producto "tarjeta" ya que los requisitos para obtener una son muy laxos. Además, la posibilidad de hacer pagos intermedios entre el mínimo y el pago total hacen a que la probabilidad de impago aumente. De acuerdo con la naturaleza propia del data set, solamente nos enfocaremos en este tipo de producto financiero.

En este sentido resulta crucial poder entender los perfiles crediticios de una institución que ofrece este tipo de servicios, ya que permitiría potencialmente una reducción de las perdidas esperadas asociadas al riesgo de crédito. En otras palabras, una mejor estimación de los perfiles crediticios *ex ante*, permitirían que las instituciones bancarias posen mejor información respecto de a quien otorgarle (y a quien no) una tarjeta de crédito.

1.3. Literatura Relacionada

La literatura de modelos que se encargan de estimar el riesgo de crédito es amplia. Se pueden encontrar dos grandes familias de modelos, contrapuestos entre sí, para medir el score crediticio: la vertiente econométrica, pone el énfasis en encontrar las variables que contribuyen al default; mientras la corriente derivada de las técnicas de minería de datos pone el énfasis en la clasificación binarias de cada observación en los dos estados posibles ("default" y "no default"). En este trabajo se hará especialmente foco en la segunda corriente ya que dichos son los que conciernen a la especialización.

En primer lugar, las primeras investigaciones donde se aplican técnicas de minería de datos a riesgo de crédito devienen de Fisher (1936) y Durand (1941). Ellos utilizan el análisis discriminatorio lineal y cuadrático respectivamente para categorizar aplicaciones de crédito entre "buenas" (el cliente tiene la capacidad de repago necesaria) y "malas" (el cliente no tiene la capacidad de repago necesaria para hacer frente a la deuda). Estos dos métodos son la base del análisis multivariante que subyacen a diferentes algoritmos clasificatorios como el principio de kernel, la regresión logística, el clasificador óptimo de Bayes, Bayes ingenuo, entre otros.

Por otro lado, Zurada & Lonial (2005) realizan una investigación donde aplican diversas técnicas de minería de datos para la clasificación del recupero de deudas incobrables en la industria de la medicina prepaga. En este caso, los autores argumentan que el incremento en el costo de la sanidad en los EE. UU. se debe al aumento en las deudas incobrables de sus pacientes. Además, dichas instituciones no suelen pedir información financiera antes de brindar su servicio, por lo que se les hace complicado saber si un paciente que se endeuda con ellas para costearlo es verosímil a repagar su deuda o no. Esto quiere decir que no hay manera de saber cuál es el score de cada cliente a la hora de ofrecer la financiación.

En dicho trabajo, los autores utilizan 5 técnicas de minería de datos para evaluar si la deuda contraída con dichas instituciones es proclive a ser repagada. Así, la red neuronal y la regresión logística, junto con su ensamble, permiten lograr la mejor precisión en la clasificación, utilizando como métrica el área bajo la curva ROC.

También, Koh et al (2006) proponen un método en dos etapas para la construcción de modelos a partir de técnicas de minería de datos. De acuerdo con esta metodología, la primera etapa conlleva la construcción del modelo individual (para ello hay que definir un objetivo, seleccionar las variables relevantes, seleccionar los datos, elegir la herramienta de modelado a utilizar, y validar el modelo. En caso de que se requiera, se puede hacer un ensamble entre distintos modelos) y luego como paso siguiente realizar una post implementación (realizar un monitoreo del modelo y reformular el modelo si es necesario). Podemos encontrar otros ejemplos en la literatura, como por ejemplo Kirkos et al (2007), donde utilizan técnicas de minería de datos para la detección de fraude en los balances de las empresas. En este caso, los resultados de utilizar 3 técnicas de minería de datos (arboles, redes neuronales, redes bayesianas) sugieren que los balances de las firmas contienen datos falsos. En términos de performance, el algoritmo que mejor logró clasificar a las firmas fue la red bayesiana (90,3% en la muestra de testeo con una validación cruzada de 10 pliegues).

De esta forma, podemos ver que la literatura relacionada con los modelos que intentan modelizar la probabilidad de default es extensa, aunque la gran mayoría se centra en las técnicas clasificatorias a partir de técnicas de minería de datos.

Revisión y análisis exploratorio del data set utilizado

La metodología utilizada en el presente trabajo es el proceso Knowledge Discovery in Databases (KDD). Su objetivo consiste buscar relaciones entre los datos que son de interés para el investigador que la vaya a utilizar. Este proceso cuenta de 5 etapas: *Integración o Selección, Preprocesamiento, Transformación, Minería de datos e Interpretación y evaluación* (Hand, 1998). En las primeras dos se seleccionan las variables y las fuentes de datos. En la segunda etapa se realiza el análisis exploratorio y la limpieza de los datos – tratamiento de valores ausentes (missing values) como de los que se encuentran fuera de rango (outliers). En la última, se generan nuevas variables que surgen a partir de los atributos originales.

Teniendo presente la naturaleza pública del data set, hay varias etapas del proceso previamente mencionado que no se llevaran a cabo. Entre ellas se omite la selección o integración del proceso KDD dado que no fue relevante "integrar" diferentes bases de datos para arribar a la finalmente utilizada. Además, se omite parte del preprocesamiento ya que la base de datos no poseía datos faltantes. Por último, se decidió no llevar acabo el tratamiento de anomalías.

2.1. Fuentes de datos

La elección del data set a trabajar provino de *"UCI Machine Learning repository"* que contiene información con relación a pagos defaulteados en tarjetas de crédito taiwanesas para 30.000 clientes, entre los meses de abril a septiembre de 2005. Inicialmente, este data set fue utilizado por Yeh & Lien (2009) para estimar la precisión predictiva de la probabilidad de default entre 6 algoritmos de minería de datos diferentes (Arboles decisorios, K-NN, regresión logística, random forest, redes neuronales, entre otros).

En el año 2005, Taiwán enfrento una crisis de deudas impagas en tarjetas de crédito alcanzando su pico máximo durante el tercer trimestre de 2006 (Chou, 2006). Entre los factores que contribuyeron a la crisis fue una sobre expansión de tarjetas por parte de instituciones financieras a personas que no calificaban para ellas. Asimismo, la mayoría de los que poseían dichas tarjetas las sobre utilizaron financiando más consumo del que podían

hacer. Esto generó que el crecimiento de las deudas, en conjunción de la irresponsabilidad en el otorgamiento de las tarjetas, generase una crisis financiera de elevada magnitud presentando grandes desafíos para ambas partes (Yeh. y Lien, 2009).

Desde el punto de la administración del riesgo, los autores argumentan que la precisión predictiva de la probabilidad es más relevante que el resultado de la clasificación predictiva. De esta forma, el resultado de la predicción de una red neuronal, entre los 6 algoritmos probados, fue aquella que permitió estimar correctamente la verdadera probabilidad de default.

Este data set contiene información socioeconómica para cada uno de los 30.000 clientes de tarjetas de crédito que va desde pagos defaulteados, factores demográficos, historial crediticio, historial de pagos, montos de tarjeta adeudados por cliente, etc. En particular, cada atributo del *data set* viene dado por:

- **ID:** es el identificador de cada cliente. En este caso, tenemos 30.000 IDs.
- **LIMIT_BAL**: es el monto de crédito otorgado a cada cliente en dólares taiwaneses.
- **SEX**: es la variable que describe el *sexo* del cliente. Toma valor 1 para masculino y 2 para femenino.
- **EDUCATION:** nivel educativo que alcanzó ese cliente. Toma valor de 1 a 6, donde 1 representa el máximo nivel de educación alcanzado, *posgrado*[1], 2 universitario completo[2], 3 *secundario completo*, 4 *otros*, 5 y 6 *desconocido* (Tabla 3).
  - **MARRIAGE:** representa si el cliente se encuentra casado (toma valor 1), soltero (toma valor 2) o tuvo otro tipo de arreglo conyugal (toma valor 3).
  - **PAY_0:** representa el *estado de repago de la tarjeta*(si vino repagando sus deudas de tarjeta de crédito o el **historial de pago del cliente**) en <u>*septiembre*</u> 2005 (-1=viene pagando a término, 1=el pago de la deuda se encontró atrasado 1 mes, 2= el pago de la deuda se encuentra atrasado 2 meses,..., 8= el pago de la deuda se encontró atrasado 8 meses, 9= el pago de la deuda se encontró atrasado 9 meses y así).
  - **PAY_2:** representa el *estado de repago de la tarjeta* (si vino repagando sus deudas de tarjeta de crédito) en <u>*agosto*</u>2005 (misma escala que para PAY_0).

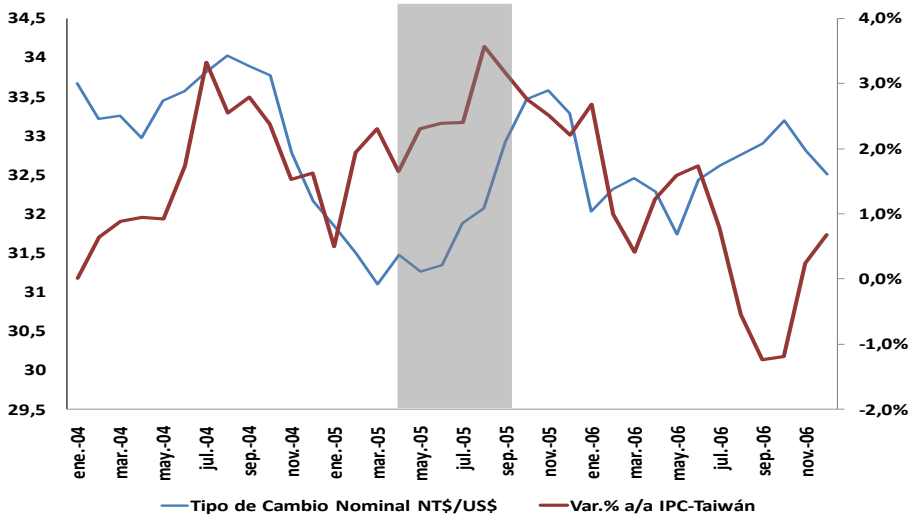
[1] El descriptor de cada variable aclara a 1 como *graduate school*, que en este caso se asume su equivalente para simplificar como *"posgrado"*.
[2]El descriptor de cada variable aclara a 2 como *university*, que en este caso se asume su equivalente para simplificar como "universitario completo".

- **PAY_3:** representa el *estado de repago de la tarjeta* (si vino repagando sus deudas de tarjeta de crédito) en *julio*2005 (misma escala que para PAY_0).
- **PAY_4:** representa el *estado de repago de la tarjeta* (si vino repagando sus deudas de tarjeta de crédito) en *junio*2005 (misma escala que para PAY_0).
- **PAY_5:** representa el *estado de repago de la tarjeta* (si vino repagando sus deudas de tarjeta de crédito) en *mayo*2005 (misma escala que para PAY_0).
- **PAY_6:** representa el *estado de repago de la tarjeta (*si vino repagando sus deudas de tarjeta de crédito) en *abril* 2005 (misma escala que para PAY_0).
- **BILL_AMT1:** monto que tuvo que pagar cada cliente de tarjeta en *septiembre* 2005.
- **BILL_AMT2:** monto que tuvo que pagar cada cliente de tarjeta en *agosto* 2005.
- **BILL_AMT3:** monto que tuvo que pagar cada cliente de tarjeta en *julio* 2005.
- **BILL_AMT4:** monto que tuvo que pagar cada cliente de tarjeta en *junio* 2005.
- **BILL_AMT5:** monto que tuvo que pagar cada cliente de tarjeta en *mayo* 2005.
- **BILL_AMT6:** monto que tuvo que pagar cada cliente de tarjeta en *abril* 2005.
- **PAY_AMT1:** monto previo pagado en *septiembre* 2005.
- **PAY_AMT2:** monto previo pagado en *agosto* 2005.
- **PAY_AMT3:** monto previo pagado en *julio* 2005.
- **PAY_AMT4:** monto previo pagado en *junio* 2005.
- **PAY_AMT5:** monto previo pagado en *mayo* 2005.
- **PAY_AMT6:** monto previo pagado en abril 2005.
- **default.payment.next.month**: default del pago del mes próximo (octubre 2005). Toma valor 1 si *defaultea*, 0 cuando no. Es la variable que se quiere predecir.

2.2. Análisis exploratorio sobre las variables continuas

Durante el período bajo estudio, Taiwán gozó de una relativa estabilidad macroeconómica. De hecho, tanto el tipo de cambio (NT$/USD) como la inflación se movieron en el mismo sentido y con órdenes de magnitud similares. Observando la banda gris en el gráfico 1, la inflación acumula 2,9% mientras que la depreciación del tipo de cambio fue del 4,9%. De esta forma, no se consideró la volatilidad macroeconómica como fuente de variabilidad que pudiese llegar a tener un impacto en la probabilidad de default, por lo que las variables expresadas en NT$ no fueron ni actualizadas por inflación ni tampoco convertidas a dólares americanos.

Gráfico 1: Inflación (Eje derecho) y Tipo de cambio en Taiwán contra el dólar estadounidense (Eje izquierdo).



Fuente: elaboración propia en base a datos de FRED y National Statistics - Republic of China (Taiwán)

En efecto, si consideramos un período más extenso de tiempo, como por ejemplo enero 2004 / diciembre 2006, la inflación y la devaluación promediaron 1,5% y 32,7 NT$ por dólar estadounidense, con desvíos estándar de 1,2% y 0,82 NT$ por dólar. Teniendo en cuenta que según la teoría económica estas variaciones son muy pequeñas para afectar las expectativas de los agentes económicos, incorporarlas como predictoras en dicho modelo no traerían aparejado una mejora de su poder predictivo (Blanchard, 2012).

En cuanto a las variables continuas, las principales estadísticas descriptivas de LIMIT_BAL y AGE nos indican que su distribución de probabilidad aproxima a una normal, aunque, en ambos casos, un poco corrida a la derecha y levemente leptocúrtica (especialmente en LIMIT_BAL). Además, se observa como el valor del límite promedio otorgado a cada cliente es de NT$ 167.000, aunque el desvió estándar es considerable (NT$ 129.748). Esto se debe al rango de variabilidad de los montos otorgados a cada cliente (desde lo NT$ 10.000 hasta los NT$ 1.000.000). Algo similar ocurre con la variable AGE, cuyo rango de variabilidad oscila entre los 21 y 79 años para la muestra de clientes: la edad promedio es de 35, pero el desvío estándar es de 9 (Tabla 1).

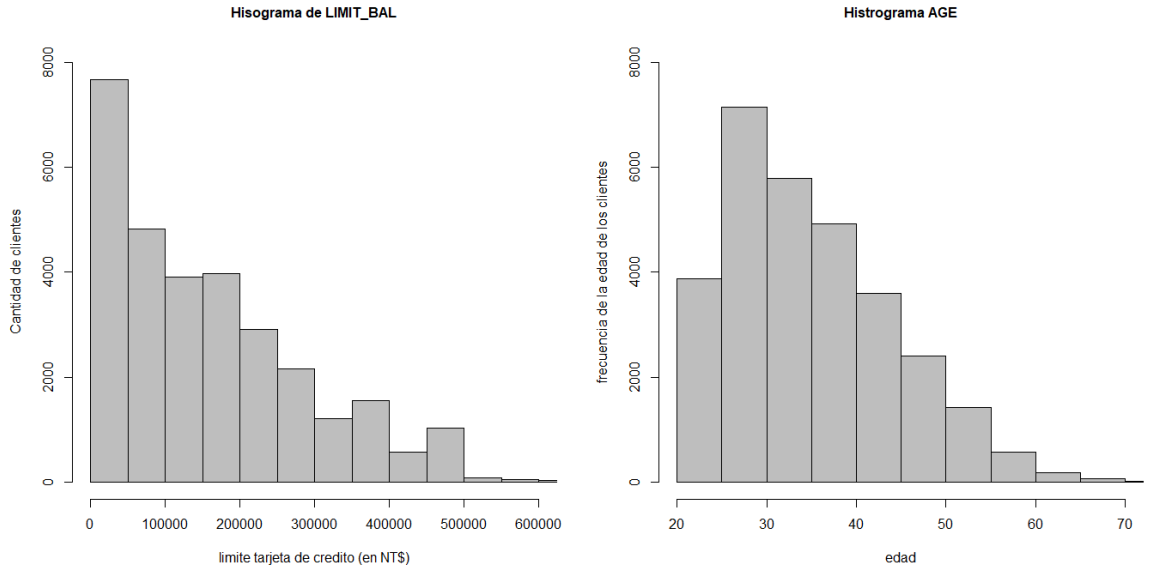
Tabla 1: Estadísticas descriptivas de LIMIT_BAL y AGE

	LIMIT_BAL	AGE
Máximo	1000000	79
Mínimo	10000	21
Promedio	167484	35
Desvío Estándar	129748	9
Percentil 75	240000	41
Percentil 50 (Mediana)	140000	34
Percentil 25	50000	28
Moda	50000	29
Curtosis	0.54	0.04
Coeficiente de Asimetría	0.99	0.73
Coeficiente de Variación	0.77	0.26

Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

Si miramos los histogramas para ambas variables podemos ver como el límite otorgado a la tarjeta de crédito, LIMIT_BAL, posee una relación negativa con la frecuencia de clientes: mayor cantidad de clientes acceden a un límite de la tarjeta “bajo” en NT$, mientras que solamente pocos clientes acceden a límites de tarjeta de crédito elevado. Por otro lado, la mayor cantidad de clientes se encuentran entre el rango etario de los 10 a 40 años. Esto nos indica que la gran mayoría de la cartera de clientes de la institución financiera se encuentra en edad activa o edad de trabajar (Gráfico 2).

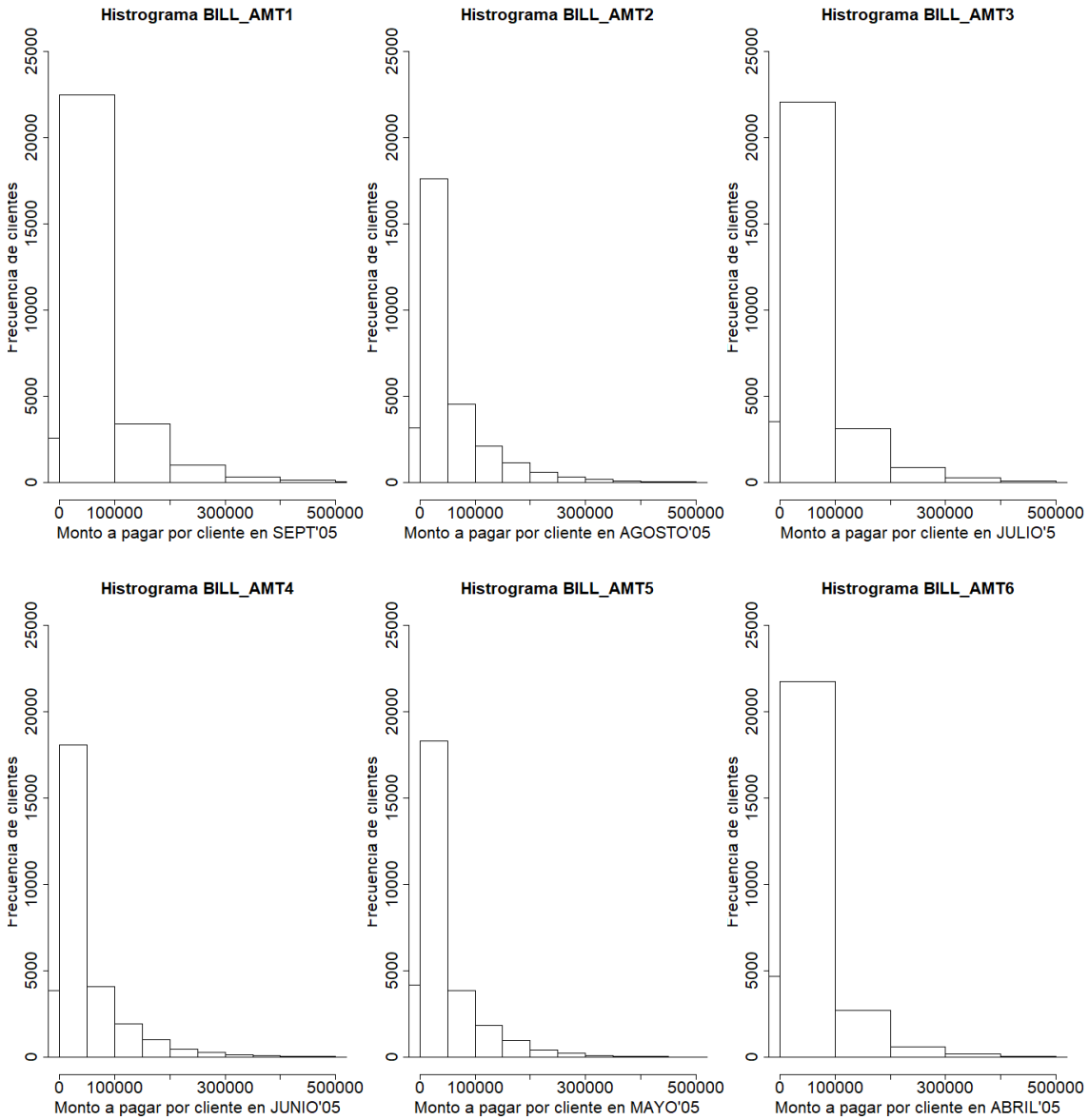
Gráfico 2: Histogramas de LIMIT_BAL y AGE



Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

El monto total por pagar (el valor de la deuda) de tarjeta de crédito para la mayoría de los clientes ($BILL\_AMT_t \; \forall \; t = 1 \ldots 6$) se concentró entre los rangos NT$ 0 – NT$ 50.000 y NT$ 50.000– NT$ 100.000 (Gráfico 3). En particular, para los meses de abril, julio y septiembre aproximadamente 22.500 clientes (75% del total de la muestra) debían pagar un monto de tarjeta de crédito entre NT$ 0 – NT$ 100.000, mientras que para los meses mayo, junio y agosto aproximadamente 17.500 a 18.000 clientes (58,33% a 60% del total) debían pagar un monto de tarjeta de crédito entre NT$ 0 – NT$ 50.000.

Gráfico 3: Histograma para $BILL\_AMT_t \; \forall \; t = 1 \ldots 6$



Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

Este tipo de análisis nos permite observar la potencial presencia de anomalías – o *outliers* como se las conoce en la jerga – debido a la elevada concentración de los montos adeudados entre 0 y 100.000 NT$ para cada uno de los respectivos meses. Si bien la corrección de este fenómeno permitiría potencialmente mejorar los resultados de las estimaciones, dicho proceso no fue llevado a cabo dado que no hace al foco del trabajo.

2.3. Análisis exploratorios sobre las variables discretas

En cuanto a las variables discretas – o categóricas – podemos decir que la base de datos posee más clientes mujeres (18.112) que hombres (11.888) (Tabla 2). Es decir, de la muestra de clientes de tarjeta de crédito un poco más de 60 % son mujeres y el resto son hombres.

Tabla 2: Clientes de tarjeta de crédito según sexo

SEX	Clientes	En %
1 (Hombre)	11888	39.6%
2 (Mujer)	18112	60.4%
Total general	30000	100.0%

Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

Por su parte, analizando la variable "EDUCATION", casi la mitad de los clientes terminó la universidad (4,8%), el 35,3% incluso tiene un posgrado realizado, y casi el resto posee título secundario o no lo informa (Tabla 3).

Tabla 3: nivel educación alcanzado de los clientes de la muestra.

EDUCATION	Clientes	En %
1 (Posgrado)	10585	35.3%
2 (Univ. Completo)	14030	46.8%
3 (Secundario)	4917	16.4%
4 (Otro)	123	0.4%
5 (Otro)	280	0.9%
6 (Otro)	65	0.2%
Total general	30000	100.0%

Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

Por su parte, más de la mitad de los individuos se encuentran solteros (53,2%) mientras que solo 1,3% de la muestra declaró otro estado civil.

Tabla 4:

MARRIAGE	Clientes	En %
1 (Casado)	13659	45.5%
2 (Soltero)	15964	53.2%
3 (otro)	377	1.3%
Total general	30000	100.0%

Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

Por último, mirando la variable dependiente "DEFAULT", la proporción de clientes que no defaultearon su tarjeta de crédito al próximo mes es el 22,1%. En otras palabras, de los 30.000 clientes, 6.636 defaultearon su tarjeta de crédito en octubre de 2005.

Tabla 5: cantidad de clientes que efectivamente defaultearon

Default Payment Next Month ?	Clientes	En %
Si	6636	22.1%
No	23364	77.9%
Total General	30000	100.0%

Fuente: elaboración propia en base a datos de Credit Card Default Risk Taiwan 2005

En efecto, se puede observar el elevado desbalance en la variable dependiente lo que podría tender a los algoritmos a clasificar los clientes como no defaulteadores cuando en realidad si lo son. Para ello, se podría utilizar algún operador para rebalancear e intentar mejorar la métrica seleccionada, aunque por falta de capacidad operativa, este tipo de análisis no pudo ser llevado a cabo.

Modelos clasificatorios para estimar el grado de default

La minería de datos puede ser definida como *la aplicación de técnicas para el análisis de datos que permita descubrir un algoritmo que producto una particular aplicación de patrones a partir de los datos.*

En este capítulo se analizarán los diferentes algoritmos y operadores, y métricas seleccionadas para poder llevar a cabo las dos últimas etapas del proceso KDD, mencionado en el apartado anterior: Interpretación y Evaluación. En ellas, se involucran las medidas de evaluación y la trasposición de resultados técnicos a niveles comerciales, de tal manera, que la aplicación del procedimiento converja a acciones correctivas en el negocio, que solucionen el fenómeno estudiado (Hand, 1998).

3.1. Algoritmos de minería de datos y métricas de evaluación utilizadas

La metodología aplicada en este trabajo fue similar a la planteada por Yeh y Lien (2009) a esta misma base de datos. En ella, los autores utilizan 6 algoritmos de aprendizaje automático para decidir sobre cuál es el mejor para determinar la verdadera probabilidad de default. Sin embargo, y debido a la capacidad de procesamiento del computador utilizado, solamente se probaron 3 algoritmos de los 6 y se incorporó el de *Random Forest*. Cada uno de ellos se caracteriza por:

Por lo tanto, los algoritmos probados fueron:

- **Random Forest**: Este concepto fue presentado por primera vez por Leo Breiman y Adele Cutler. El mismo es un tipo de ensamble que utiliza una técnica similar al Bagging. Al decidir dividir cada nodo en un árbol de decisión, a diferencia de Bagging, Random Forest solo considera un subconjunto aleatorio de todos los atributos en el conjunto de entrenamiento. Para reducir el error de generalización, el algoritmo se aleatoriza en dos niveles, selección de registros de entrenamiento y selección de atributos, en el funcionamiento interno de cada clasificador base (Kotu & Deshpande, 2014).
- **Decision Tree:** es un algoritmo de clasificación que se utiliza para separar un conjunto de datos en clases que pertenecen a la variable de respuesta. Por lo

general, la variable de respuesta tiene dos clases: Sí o No (1 o 0). Si la variable de respuesta tiene más de dos categorías, entonces se han desarrollado variantes del algoritmo del árbol de decisión que pueden aplicarse. En cualquier caso, los árboles de decisión se usan cuando la respuesta o la variable objetivo es de naturaleza categórica (Quinlan, 1986).

- **K-NN**: es un método no paramétrico, no asume distribuciones subyacentes en los datos, sólo asume que inputs similares tendrán outputs similares. La k en el algoritmo k-NN indica el número de registros de entrenamiento cercanos que deben considerarse al hacer la predicción para un registro de prueba sin etiqueta. Como la clase del registro objetivo se evalúa mediante votación, a k generalmente se le asigna un número impar para un problema de dos clases (Kotu & Deshpande, 2014). Los autores destacan que el algoritmo maneja de manera natural los problemas multiclase.
- **Regresión Logística:** es un método que extiende la idea de regresión lineal a la situación donde la variable dependiente, *Y,* es categórica. Normalmente, la variable categórica puede tomar dos estados, aunque puede existir casos con más aún. Para esta base da datos, la variable dependiente toma solo dos estados posibles – default o no default – representados por unos (1) y ceros (0).

El criterio elegido para determinar que tan bien clasifica los distintitos tipos de algoritmos a cada uno de los clientes fue el área bajo la curva ROC. Dicha métrica sirve para representar el ratio de verdaderos positivos – aquellos que el modelo clasifica como morosos cuando en el data set también lo son – frente a la razón de falsos positivos – aquellos individuos clasificados como morosos cuando, y de acuerdo con el set de datos, estos no lo son.

Debido a que la base de datos posee una mayor proporción de clientes que no defaultearan que aquellos que si (77,9% vs 22,1% respectivamente), los resultados indican que todos los modelos tienden a clasificar mejor a los primeros que los segundos. En definitiva, el margen de error es insensible a la exactitud (*accuracy*) de la clasificación en cada caso. Por lo tanto, el área bajo la curva ROC (AUC-ROC) nos da una mejor solución para comparar la performance de los distintos modelos (Yeh & Lien, 2009).

3.2.Aplicación del modelo a través de técnicas de aprendizaje automático

En el presente trabajo, las técnicas de minería de datos utilizadas tienen por objetivo lograr la clasificación binaria de clientes de acuerdo con su probabilidad de default. En este sentido, la variable dependiente toma valor 1 o 0 en función si el individuo defaulteó sus pagos con tarjeta de crédito al mes siguiente. Esta variable va a intentar predecirse a partir de todo el conjunto de variables predictoras previamente mencionadas en el capítulo anterior.

Si bien Yeh & Lien (2009), los autores que utilizan por primera vez este data set separan la base en entrenamiento y validación de forma estratificada, en este trabajo se propuso realizar una validación cruzada de 10 folds y muestreo estratificado. Esto se realizó para saber cuál es la amplitud del intervalo de confianza de cada curva ROC y evitar el sobreajuste (overfitting) de los datos por parte de los modelos.

Otros de los operadores utilizados fue utilizar "bagging" para ensamblar los modelos y mejorar la clasificación. El sample ratio fue de 0,85 y la cantidad de iteraciones de 10. También, se llevó a cabo una normalización de las variables a partir de la transformación Z y las variables fueron llevadas de nominales a numéricas.

El conjunto de datos descrito fue levantado por el software de libre uso R-Studio el cual permitió llevar a cabo, no solo en análisis exploratorio de cada variable, sino también correr cada uno de los algoritmos previamente mencionados. Así, todos los modelos fueron corridos con los parámetros definidos por default en R, a excepción de Desicion Tree y Random Forest que se aplicó preprunning y prunning para ambos casos y la profundidad máxima (maximal depth) fue de 10 para ambos.

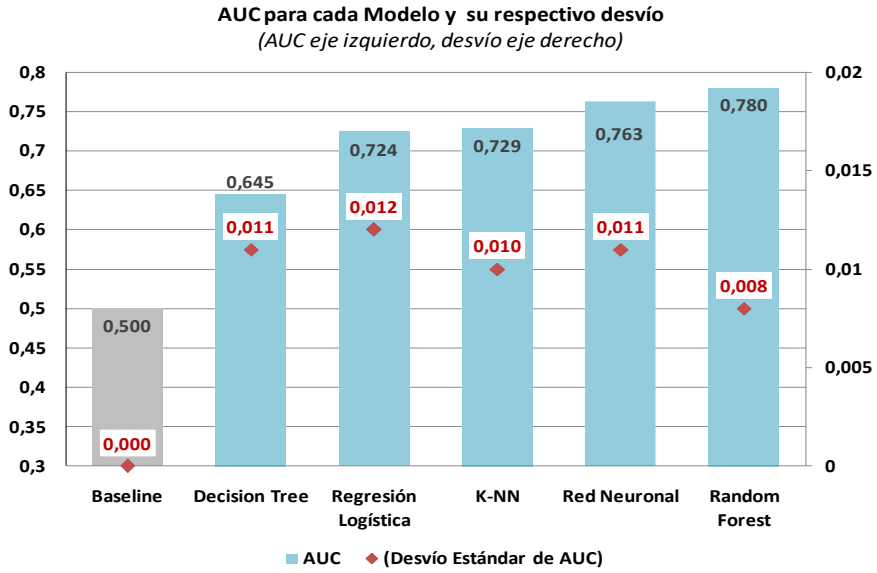
Además, a partir de probar cada algoritmo bajo un mismo conjunto de operadores y quedarnos con aquel que brinde un mejor AUC-ROC (es decir, que clasifica mejor a cada cliente en default y no default), se procedió a trabajar sobre dicho aplicando diversas técnicas que mejorasen la métrica elegida.

3.3. Resultados obtenidos e interpretación de resultados

A partir de aplicar los distintos modelos podemos concluir que el mejor algoritmo para clasificar los diferentes modelos es Random Forest, tanto porque el área bajo la curva –AUC- es el mayor como así también su desvío es el más bajo. En este caso, el algoritmo de

Random Forest mejora la clasificación de cada observación en 0,28 respecto al caso en que sean clasificados por "azar" (baseline: 0,5) (Gráfico 4).

Gráfico 4: AUC y sus respectivos desvíos en cada caso.



Fuente: elaboración propia en base a datos de *Credit Card Default Risk Taiwan 2005*

Además, este algoritmo no solo es el mejor en cuanto que maximiza la métrica AUC ROC, sino que además brinda el menor desvío con relación al resto. En este caso, el desvío estándar del AUC ROC es 0.008, siendo muy inferior al de la regresión logística (0,012).

Así, se obtuvo una mejora del 56% respecto del escenario base lo cual implica una ganancia tanto para las instituciones financieras desde el punto de vista del negocio en sí mismo. Además, representa una herramienta muy útil para poder discernir con una mejor exactitud entre los diferentes tipos de clientes – aquellos con más y con menos capacidad de repago – para tomar la decisión de a quien prestar y a quien no.

Conclusiones y Futuros Estudios

4.1. Conclusiones

El objetivo del presente trabajo era intentar responder a la siguiente pregunta: ¿Cuáles técnicas de minería de datos permiten realizar una mejor clasificación binaria de clientes en tarjetas de crédito para determinar si defaultearán o no sus pagos al mes siguiente, dada una cierta confianza que eso sea cierto?

En efecto, a lo largo de todo el trabajo se intentó alcanzar una conclusión que responda a la pregunta planteada y, que, a su vez, fuese de utilidad para cualquier tipo de entidad financiera donde la necesidad de conocer la probabilidad de que un cliente no pueda hacer frente a sus deudas sea una fuente de información *clave* para poder llevar acabo dicho negocio.

Para ello, primero se estableció el marco teórico/conceptual que fue utilizado a lo largo de todo el trabajo. En él, se brindaron las principales definiciones para tener en cuenta para poder poner al lector en sintonía con lo expuesto en capítulos posteriores y acotar el campo de estudio a unas pocas cuestiones dentro de esta problemática. Por ejemplo, se estableció que se consideraba por riesgo de crédito – y que no – y cuales fueron los sinónimos adoptados para esta definición.

Además, se utilizó un enfoque *deductivo* partiendo de cuestiones muy generales (que es el "Big Data", como puede ser definido, etc.) a cuestiones más particulares (su campo de aplicación; el "text mining" y las restricciones legales para aplicar dichas técnicas al riesgo crediticio, etc.).

Luego, se realizó un breve racconto de la literatura asociada con modelos que intentan estimar la probabilidad de default, definiendo dos tipos: la vertiente econométrica (que pone énfasis en encontrar las variables que contribuyen al default) y la vertiente asociada a las técnicas de minería de datos (que pone énfasis en la clasificación de cada cliente en dos posibles estados "default y no default".

Se aplicó la metodología KDD – *Knowledge Discovery Databases* – para buscar relaciones entre los datos que son de interés para el investigador que la vaya a utilizar. Así se

realizó un análisis exploratorio de cada variable del data set, separando entre aquellas que son continuas (como la edad, el límite de la tarjeta de crédito y los montos adeudados en cada momento del tiempo, etc.) y aquellas que son discretas (el sexo, el estado civil, cantidad de cuotas atrasadas, etc.).

Por último, se definió la metodología utilizada para estimar la probabilidad de default en cada caso. Para ello, primero se definió la métrica para evaluar la bondad de dicha probabilidad estimada – área bajo la curva ROC. Luego, se definieron los diferentes operadores bajo los cuales fueron probados cada modelo – cross validation con 10 folds y bagging, normalización Z de cada variable, etc. Por último, fueron definidos cada uno de los algoritmos testeados para llegar a la conclusión de que, *bajo la configuración de parámetros y operadores seleccionados, random forest es aquel que maximiza el área bajo la curva ROC.*

En otras palabras, para la base de datos tomada como ejemplo, de los cinco algoritmos de clasificación entrenados (Decision Tree, K-NN, Random Forest, Redes Neuronales y Regresión Logística), *Random Forest* es el que maximiza el área de la curva ROC, siendo el que, brinda el menor desvío con relación al resto.

4.2. Futuros estudios

Teniendo en cuenta los resultados obtenidos, futuros trabajos requerirán probar diferentes técnicas para mejorar la métrica seleccionada. Las estadísticas descriptivas evidencian la concentración del monto adeudado para la gran mayoría de los clientes entre NT$ 0 / 50.000 y NT$ 50.000 y NT$ 100.000 lo cual podría ser utilizado para “agrupar” *(clusterizar)* clientes en distintos grupos.

En línea con lo anterior, podría ser necesario llevar a cabo un análisis de anomalías – o *outliers* como se las conoce en la jerga – debido a la elevada concentración de los montos adeudados entre 0 y 100.000 NT$ para cada uno de los respectivos meses. De esta forma, la corrección de este fenómeno permitiría potencialmente mejorar los resultados de las estimaciones.

También, teniendo presente que la base se encuentra desbalanceada en favor a aquellos que no defaultearon en octubre 2005 (22,1% de clientes que defaultearon y 79,1% que no) se podría utilizar algún operador para rebalancear e intentar mejorar la métrica seleccionada.

Por último, no debemos descartar la creación de nuevos campos que permitan nutrir al modelo de información relevante para la clasificación. Por ejemplo, se podrían crear variables dummies para categorizar a los clientes de acuerdo con los montos que adeudan y nutrir al modelo de información relevante para la clasificación. En línea con esto, no habría que dejar de modelizar los comportamientos en determinadas variables macroeconómicas que podrían llegar a tener impacto en el repago (o no) de la tarjeta de crédito para los clientes de la muestra.

Referencias bibliográficas

Akerlof G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. The Quarterly Journal of Economics, Vol. 84, No. 3 (Aug. 1970), pp. 488-500. The MIT Press.

Byte Ti, R. (2017, diciembre 26). Big Data en el Sector Financiero. Recuperado de https://revistabyte.es/actualidad-byte/big-data-sector-financiero-2/

Chang, C. H.; Chang, H. H.; & Tien, J.-C. A Study on the Coping Strategy of Financial Supervisory Organization under Information Asymmetry: Case Study of Taiwan's Credit Card Market (2017). Universal Journal of Management 5(9): 429-436, 2017

Chou, M. (2006). Cash and credit card crisis in Taiwan. Business Weekly, 24–27.

Chye Koh, H., Chin Tan, W., and Peng Goh, C., 2006, "A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques." Journal of Business and Information, 1, 96-118.

Douglas, L. (2011). 3d data management: Controlling data volume, velocity and variety. Gartner. Retrieved.

Durand, D., 1941, Risk elements in consumer installments financing.

Fisher, R., 1936, "The use of multiple measurements in taxonomic problems." Annals of Eugenics, 7, 179–188

Galindo, J., and Tamayo, P., 2000, "Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications." Basic Methodology and Risk Modeling Applications 15(1-2), 107–143.

HAND, D. J. (1998). “Data mining: statistics and more?”. The American Statistician, 52(2), 112-118.

Itgovernance.co.uk. 2016. The Basel Accords - An Overview of The Basel Financial Regulations: Basel I, Basel II And Basel III. [online] Available at: <http://www.itgovernance.co.uk/basel.aspx> [Accessed 25 April 2016].

Kirkos, E., Spathis, C., and Manolopoulos., Y., 2007, "Data Mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32(4), 995-1003.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety (). META Group.

Munafo, F (2019). La importancia de la gestión de datos y su impacto en el riesgo de crédito de instituciones financieras. Revista de investigación en modelos financieros año 8 volumen II (2019-II)

Republica del Ecuador (2003). Superintendencia de Bancos Y Seguros. Libro I, Normas Generales para las Instituciones del Sistema Financiero; Titulo X, Capítulo II, Sección I.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., &LichtendahlJr, K. C. (2018). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

Zurada, J., and Lonial, S., 2005, "Comparison of the Performance of Several Data Mining Methods for Bad Debt Recovery in the Healthcare Industry." the Journal of Applied Business Research, 21(2), 37-53.