Estimación de probabilidad de no pago en nuevas originaciones de tarjeta de crédito

Autores[1]: Alejandra Polanco Rojas; Mauricio De Jesús Rúgeles Vásquez; Augusto Umaña Ruiz
Tutor interno[2]: Sergio Andrés Cabrales
Tutor Externo[3]: Edgar Escobar

Resumen.

La colocación de créditos es una de las principales fuentes de ingresos de las entidades financieras. Para controlar la exposición al riesgo es necesario identificar los clientes que tienen una baja probabilidad de pagar y de esta forma tomar decisiones en función de la aversión al riesgo de la entidad. Para predecir los clientes que no pagan y los que sí, las entidades financieras emplean modelos predictivos construidos por las centrales de riesgo y usan la información de hábito de pago de los clientes que tiene cada entidad. En este estudio se plantean cuatro modelos de calificación de riesgo de crédito: *Support Vector Machines*, *Boosting*, *Random Forest* y Regresión Logística. La base de datos con la que se trabajó fue proveída por la compañía Experian y corresponde al Bureau de crédito de un país centro americano. Los resultados de este trabajo muestran que *Boosting* es la mejor estrategia para calibrar los modelos de clasificación de riesgo de crédito y que *Random Forest* es el modelo que mayor utilidad esperada.

Palabras Clave: Credit Scoring; Machine Learning; Support Vector Machines; Regresión Logística; Boosting; Random Forest; tarjeta de crédito; originación de crédito.

1. Introducción.

El negocio de banca masiva actualmente requiere de procesos que permitan aprobar créditos de forma ágil al menor costo posible y dentro de los niveles de riesgo estipulados por la dirección de cada compañía (bancos, compañías de financiamiento, supermercados, etc.). En la actualidad las instituciones financieras emplean los llamados *credit scoring models* para medir el nivel de riesgo de crédito de sus clientes y decidir si es viable dentro de sus políticas de riesgo otorgarle un crédito (Anderson, 2007).

Aunque dichos modelos de *Scoring* se han desarrollado desde la década de 1960 (Altman, 1968). La metodología predominante para calcular el riesgo de no pago de los clientes ha sido el uso de modelos de regresión logística. La

[1] a.polanco@uniandes.edu.co, md.rugeles@uniandes.edu.co, a.umana@uniandes.edu.co: Estudiantes de Maestría en Inteligencia Analítica Para la Toma de Decisiones, departamento de Ingeniería Industrial, Universidad de los Andes, Bogotá Colombia.
[2] s-cabral@uniandes.edu.co Profesor Visitante; departamento de Ingeniería Industrial, Universidad de los Andes.
[3] Vicepresidente Analytics BD; Experian (En Colombia conocida como Data Crédito es la Central de Información Crediticia líder en el mercado andino., que provee soluciones integrales a los principales sectores de la economía para la toma de mejores decisiones en el ciclo de otorgamiento de crédito).

razón de su dominancia radica en que es una metodología ampliamente conocida, relativamente fácil de implementar e interpretar en los sistemas de crédito. Por sus características computacionales, el desarrollo de estos modelos es más rápido que con metodologías modernas como los *Support Vector Machines (SVM)* o *Random Forest.*

Sin embargo, dados los avances en *Machine Learning* y la gran capacidad de cómputo que actualmente se puede conseguir a bajo costo, cabe la pregunta de si el uso de metodologías modernas como *Machine Learning* puede mejorar la capacidad de predicción de los modelos de riesgo de crédito y de esta forma aumentar la utilidad de las compañías.

En este trabajo se emplearon los modelos de *Machine Learning: SVM*, *Boosting*, *Random Forest* y Regresión Logística. Con el fin de estimar la probabilidad de no pago en nuevas originaciones y evaluándolos por su potencial para maximizar la utilidad de las compañías que los llegarán a usar para otorgar créditos.

La base de datos suministrada por la compañía Experian contiene el estado de las obligaciones reportadas al bureau de un país centro americano entre marzo de 2010 y mayo de 2014.

2. Contexto de negocio

Durante estos cuatro años la cantidad de obligaciones reportadas creció un poco más del 45%, al pasar de 5 a 7,3 millones. Sin embargo, llama la atención que el saldo reportado creció casi un 102% al pasar de US$21.3 mil millones a US$42.8 mil millones, como se puede observar en la **¡Error! No se encuentra el origen de la referencia.**.

Al abrir los datos por sector económico se encontró que el sector financiero representa el 54% del crecimiento en obligaciones reportadas y el 99% del crecimiento en saldo (Figura 2).

El crecimiento de las obligaciones estuvo concentrado en Tarjetas de Crédito y en Crédito Personal, mientras que, como lo muestra la Figura 3, en el saldo los mayores crecimientos se dieron en Préstamo Comercial y en Crédito Hipotecario.

No obstante, el crecimiento en saldo, la calidad de cartera (medida como cartera en mora mayor a 30 días/cartera total), de los Créditos Hipotecarios y Comerciales no presentó mayores deterioros en el período analizado. En la Figura 4 se ve como la cartera de Tarjeta de Crédito presenta valores de mora muy altos, estando por encima de 25% hasta mediados de 2013. Aunque la calidad de la cartera mejoró durante el segundo semestre de 2013, se observa que volvió a deteriorarse en 2014, revelando que tiene serios problemas de calidad.

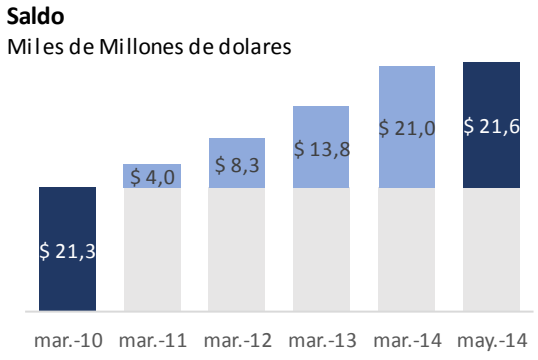

Figura 1 Saldo en miles de millones de dólares de las obligaciones.

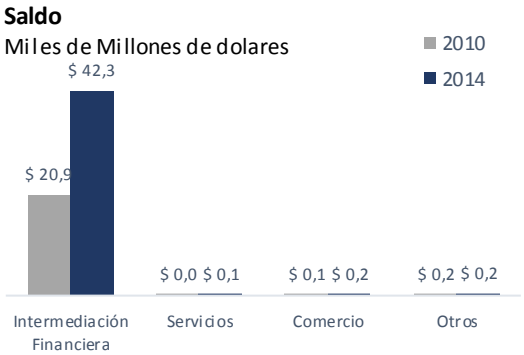

Figura 2 Saldo en miles de millones de dólares de las obligaciones por sector económico.

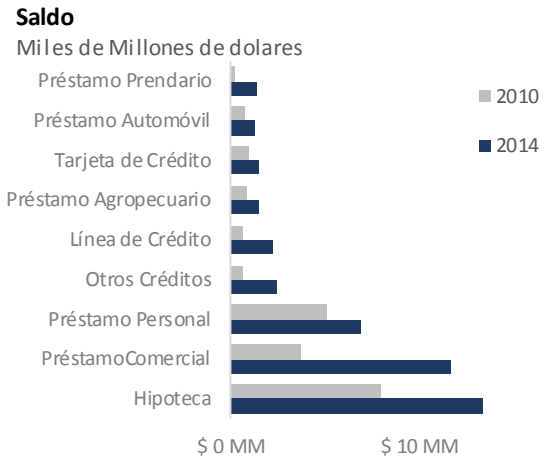

Figura 3 Saldo en miles de millones de dólares de las obligaciones del sector financiero.

Debido a este comportamiento se decidió enfocar el desarrollo de un modelo en la línea de tarjeta de crédito para predecir la probabilidad de no pago de los clientes a los que se les otorgue una nueva tarjeta de crédito.

Con este modelo de riesgo de originación, las entidades de crédito podrían mejorar la calidad de cartera de este producto, generando importantes disminuciones en su costo de crédito y por lo tanto incrementando su utilidad.

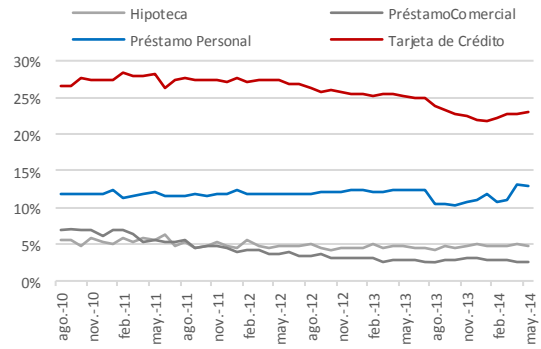

Figura 4 Comportamiento histórico de la cartera de las cuatro líneas de crédito más relevantes

3. Metodología

En este estudio se utilizaron cuatro modelos de calificación de riesgo de crédito, *Support Vector Machines*, *Boosting*, *Random Forest* y Regresión Logística.

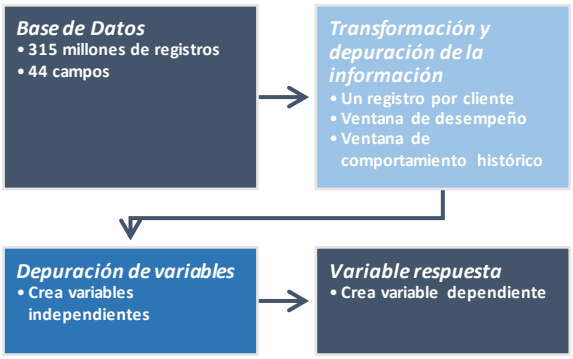

Figura 5 Diagrama de la metodología

PASO I: *Descripción de la base de datos.*

La base de datos contiene un resumen mensual de todos los productos activos e inactivos que tienen los clientes reportados en el Bureau. Está compuesta por 44 campos de información y aproximadamente 315 millones de registros. Dichos campos pueden ser clasificadas de acuerdo a su naturaleza:

a) *Identificación:* Campos tipo caracter que permiten identificar los clientes, los productos y los asociados (entidades que reportan al bureau); tipo y número de documento, número de cuenta e identificadores de los asociados. Se incluyen las fechas de reporte de la información.
b) *Obligación:* Campos tipo numérico que definen las obligaciones de los clientes; el importe o capital inicial desembolsado; la cantidad de cuotas pactadas; el importe de pago o cuota y las fechas de inicio y terminación. Se incluyen variables relacionadas con el comportamiento de la obligación tales como: saldo actual; número de días de mora; monto de los pagos realizados; vector de hábito de pago; promedio de hábito de pago y contadores de moras.
c) *Clasificación:* Campos categóricos que caracterizan las obligaciones; grupo económico de la obligación (financiero, servicios, otros); estado (activo e inactivo); tipo de producto (tarjeta de crédito, préstamo personal, crédito hipotecario, etc.) y tipo de pago.

PASO II: *Transformación y depuración de la información.*

Para la evaluación del comportamiento del cliente se dividió la base de datos en dos ventanas de tiempo. La primera se llamó ventana de desempeño y la segunda ventana comportamiento histórico.

La ventana de desempeño hace referencia al comportamiento del cliente, si incumplió o no con el pago de su tarjeta de crédito en el período de observación, este comprende los primeros doce meses de vida de la tarjeta.

Para la ventana de comportamiento histórico se concidera toda la información reportada al bureau dos años antes del otorgamiento de la tarjeta.

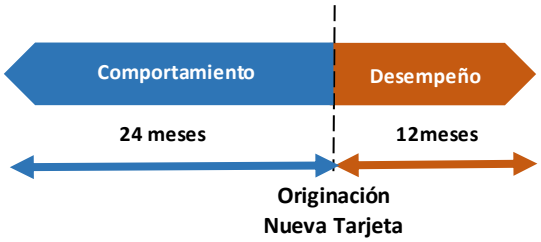

Figura 6 Ventanas de tiempo.

Dado que el conjunto de datos de estudio comprende observaciones entre marzo de 2010 y mayo de 2014, se seleccionaron las tarjetas de crédito otorgadas entre marzo de 2012 y mayo de 2013. De esta forma las observaciones cumplen con los tiempos de las ventanas de desempeño y comportamiento. Finalmente, se obtuvo una base con 102,456 clientes que cumplen estos requisitos.

Para clasificar el riesgo de crédito en las nuevas originaciones de tarjetas, se consolidó en un registro por cliente/producto, la información detallada de cada mes. Para las variables numéricas se calcularon los valores máximos, mínimos y promedios; para las variables categóricas, se realizaron conteos de los diferentes niveles. Adicionalmente se establecieron medidas de antigüedad de las obligaciones e indicadores de endeudamiento. Con esto, se resume la información de cada cliente en 330 variables al momento de solicitar una nueva tarjeta de crédito.

PASO III: *Depuración de variables.*

Debido a que las variables fueron construidas consolidando información detallada algunas de estas tienen valores ausentes (NA's) que corresponden a clientes que no tiene determinado producto y por lo tanto no se puede calcular un valor para esa variable. Estos valores ausentes generan errores al ejecutar los modelos.

Por las restricciones de los modelos de *Random Forest* y *Support Vector Machines* (SVM), se realizó el siguiente proceso de imputación de datos: si menos del 25% de los valores eran nulos, se reemplazaron con la media en las variables numéricas y con la moda en las variables categóricas. Si los valores nulos estaban entre el 25% y el 50% se categorizó la variable en tres niveles: el primer nivel "Valores ausentes", el segundo los valores inferiores a la media y el tercero los valores superiores. Si la proporción de ausentes era superior a 50% se generó una variable indicadora así: 0 si el valor corresponde a un NA, 1 en caso contrario.

Además, se identificó la presencia de valores atípicos que distorsionan la distribución de las variables independientes, por lo que fue necesario truncar los valores extremos al percentil 99 y realizar transformaciones logarítmicas a las variables numéricas.

PASO IV: *Variable respuesta.*

La variable respuesta para clasificar las obligaciones que sí se pagaron de las que no, es binaria, dónde 1 significa incumplimiento (*Default*) si un cliente alcanza una mora de más de 90 días o alcanza más de tres moras de 60 días en un periodo de observación de doce meses. Si no cumple ninguna de estas dos

condiciones, se considera un buen cliente (*No Default*). (Anderson, 2007)

4. Principales variables utilizadas.

Las principales variables utilizadas en los métodos de modelaje, se pueden resumir en:

a. *Comportamiento:* Promedio de los vectores de hábito de pago.
b. *Endeudamiento:* Razón entre los saldos promedios y los montos originalmente desembolsados.
c. *Antigüedad:* Meses de historia de los diferentes productos.
d. *Montos pagados:* Importe de los últimos pagos realizados (total y en relación con el saldo).
e. *Saldo promedio:* Saldo promedio de las obligaciones.
f. *Monto:* Capital otorgado o desembolsado en los diferentes productos.
g. *Mora:* Altura de mora máxima y contadores de altura de mora en la historia del cliente.

5. Modelos para estimar el riesgo de crédito.

Se implementaron cuatro tipos de modelos de clasificación. Se ajustó un modelo de regresión logística por ser el método clásico de modelamiento en este tipo de problemas y el empleado actualmente por la compañía Experian. Adicionalmente manifestaron su interés en evaluar el desempeño de los *Random Forest* en este tipo de problema. Se presentaron dos metodologías adicionales (*Boosting* y *Support Vector Machines*) debido al buen desempeño mostrado en artículos académicos (Cheng-Lung,2007; Tomczak, 2014) y que son los métodos que tienden a ganar las competencias de *Kaggle*[4].

El desempeño de los modelos en esta etapa se hizo mediante el AUC de la curva ROC. Los modelos escogidos se ajustaron usando una muestra aleatoria de entrenamiento con 39,059 observaciones y validaron contra una muestra de prueba de 13,020 observaciones.

Los meta-parámetros de cada modelo se calibraron usando 10-*fold Cross Validation*, por facilidad computacional, que es una práctica generalmente adaptada y además evita problemas del *trade-off* entre sesgo y varianza que generalmente se da en estos casos. (Gareth, 2015-P. 182)

Regresión logística: Con las variables creadas se calcula un modelo $logit$ con una penalización tipo $Lasso$ (con $\lambda = \text{EXP}(-10)$). A las variables resultantes del proceso $Lasso$ se les realizó un proceso *Step Wise* como criterio de selección de AIC.

Random Forest: Con las variables creadas se generó un modelo de regresión logística univariado, con el fin de identificar las variables que individualmente son significativas para explicar la relación con la variable respuesta. El modelo se ajustó con las variables que resultaron relevantes, es decir, con parámetros significativamente diferentes de cero a un nivel de confianza de $\alpha = 0.05$ y con un $AUC > 0.5$.

Adicionalmente, se realizó un algoritmo iterativo para selección de variables, basado en la importancia de estas para el modelo de *Random Forest* determinado por el índice de *Gini*. El algoritmo realiza cinco iteraciones comenzando con el número total de variables y elimina, en cada paso, la mitad de las variables menos relevantes. Para cada uno de los modelos, se encontró el mejor parámetro de *tunning* ($mtyr$), que determina la cantidad de variables que se seleccionan aleatoriamente en cada árbol, en la mayoría de los casos $mtyr = \sqrt{\#\,Variables}$.

[4] Kaggle: Sitio de internet (www.kaggle.com) especializado en administrar competencias *de Machine Learning* para resolver problemas de diversas empresas.

Boosting: Utilizando las variables depuradas a través de *Random Forest* vía *Cross Validation*, se estimaron árboles de clasificación a través de la metodología de ensamblaje *Boosting*.

Support Vector Machines (SVM): Para la calibración del modelo de SVM se usaron las variables previamente identificadas como significativas, ya que esta metodología de modelamiento no cuenta con un proceso o algoritmo de selección de variables. Se probaron cuatro tipos de *kernel*: lineal, polinomial, gaussiano (también conocidos como *Radial Basis Functions*) y laplaciano.

Debido a la complejidad computacional que tienen este tipo de modelos, la calibración se hizo mediante 10-Fold *Cross Validation* con una sub muestra aleatoria de 5,000 observaciones.

Los *kernels* lineal y polinomial tuvieron muy mal desempeño al no superar el 52% de AUC de la curva ROC. Debido a este desempeño en la etapa de calibración se descartaron como candidatos.

6. Resultados de los modelos.

Las Figura 5 y 6 muestra la comparación de los cuatro modelos por los indicadores de AUC de las curvas ROC y KS. Encontramos que el modelo que genera mejores resultados es el modelo *Boosting* que ajusta con un AUC de 77% y un KS de 41.42%, este modelo se calibró con 31 variables asociadas a saldos promedios, antigüedad promedio, montos originales y al estado de los productos financieros.

Tabla 1 Resumen Indicadores de comparación

Tipo de Modelo	Área bajo la curva ROC	KS
Modelo de regresión logística	0.763	0.3978
Modelo Random Forest	0.761	0.3918
Modelo Boosting	0.770	0.4142
Modelo Support Vector Machines	0.707	0.3157

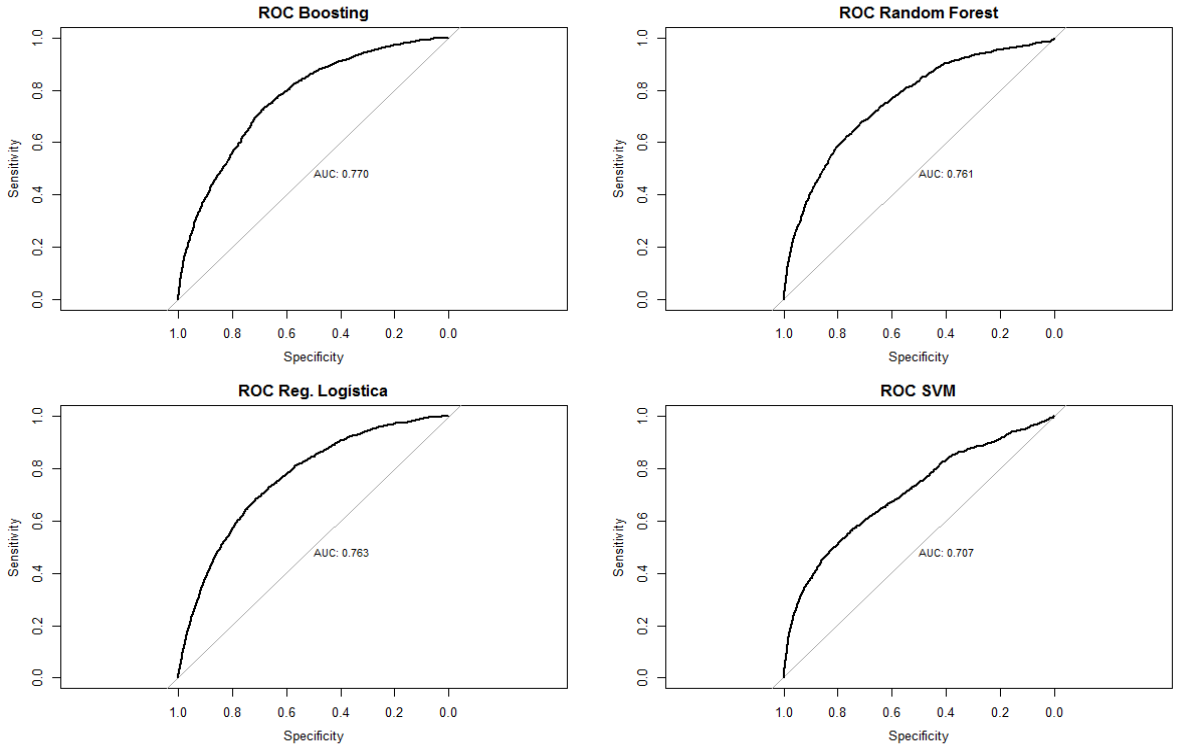

Figura 7 Criterios de AUC para los modelos seleccionados.

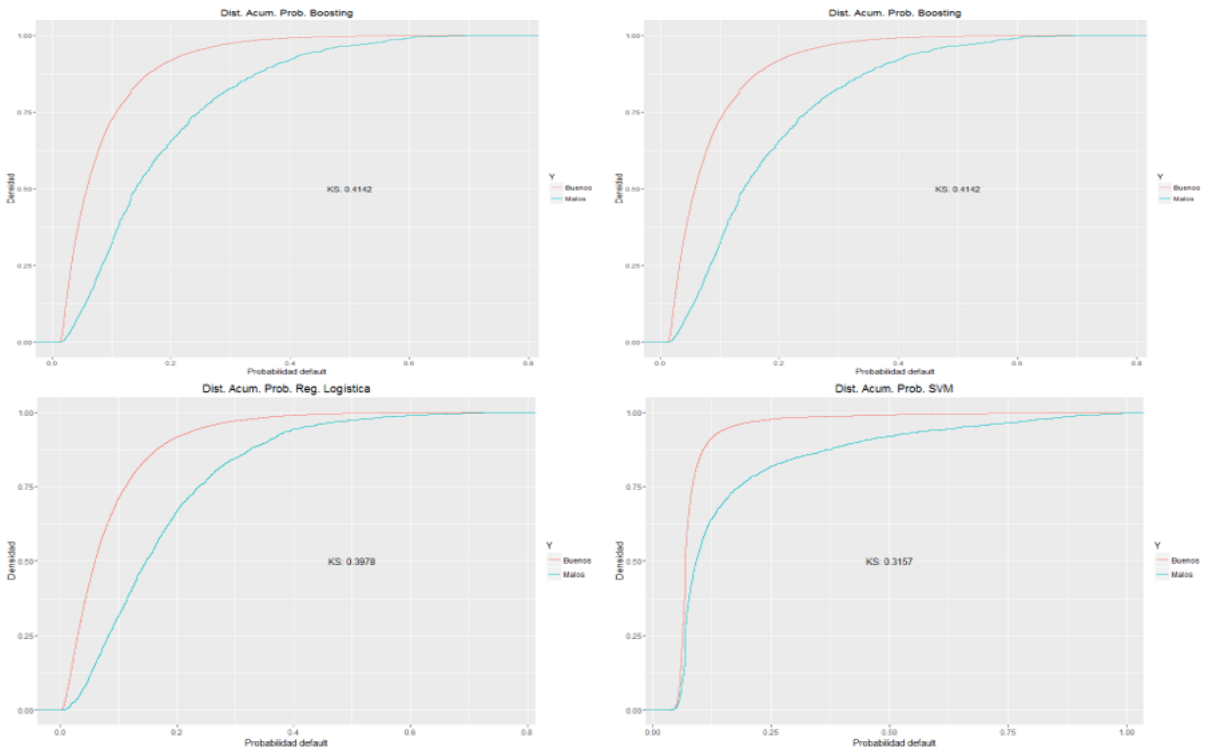

Figura 8 Criterio KS para los modelos seleccionados.

7. Función de utilidad de los modelos.

Para la compañía que otorga el crédito la capacidad predictiva de los modelos no es suficiente, ya que su objetivo es maximizar la utilidad de su portafolio a través de la correcta selección de clientes. Por eso se vio necesario crear una función de utilidad de forma que permita calcular para cada punto de corte del modelo el retorno esperado para ese nivel de riesgo.

Como se puede ver en la Tabla 2 Utilidad de la matriz de confusión, al establecer un punto de corte en un modelo de clasificación existen cuatro posibles resultados:

1. Predecir que el cliente no va a pagar (hace D*efault*) y efectivamente no paga (*Default*).
2. Predecir que el cliente no va a pagar y sí paga.
3. Predecir que el cliente sí va a pagar y en la realidad no paga.
4. Predecir que sí va a pagar y efectivamente paga.

Los casos 1 y 4 corresponden a predicciones correctas. El caso 1 es una decisión que no genera ingresos ni pérdidas, ya que esos recursos no se prestan ni generarían eventualmente algún ingreso si se prestaran. En el caso 4 el ingreso corresponde a los intereses generados por el saldo de la deuda de los clientes.

Los casos 2 y 3 son errores del modelo los cuales generan pérdida. En el caso 2 hay un costo de oportunidad, ya que esos clientes hubieran generado ingresos para la compañía. En el caso 3 se pierde el saldo que se le prestó al cliente y efectivamente no pagó.

La Ecuación 1 y la Tabla 2 resumen los cuatro escenarios explicados anteriormente.

Tabla 2 Utilidad de la matriz de confusión

		Estado Real	
		Default	**No Default**
Predicción	**Default**	0	$-r\times\sum_{N_2} S$
	No Default	$-\sum_{N_3} P$	$r\times\sum_{N_4} S$

Ecuación 1

$$Utilidad(th) = -r\times\sum_{N_2} S - \sum_{N_3} P + r\times\sum_{N_4} S$$

- P: Saldos de los créditos que no se pagan.
- S: Saldos de los créditos que se pagan.
- r: Tasa de interés.
- th: Punto de corte.

Para los modelos seleccionados, se calculó la utilidad para todos los puntos de corte usando una tasa de interés de 18% e.a. Como se muestra en la Figura 9 con el modelo *Boosting*, se tiene la utilidad más alta, aunque el modelo *Random Forest* es más estable y tiene altas utilidades con puntos de corte altos.

Llama la atención el contraste de la forma de las curvas de utilidad. En los mejores modelos según el AUC (*Boosting* y Regresión logística) la utilidad cae rápidamente y para puntos de corte de 12.5% los portafolios otorgados por modelos darían pérdida. Sin embargo, los otros dos modelos pueden tolerar niveles de punto de corte más altos aun dando utilidad.

Esto ocurre porque los modelos de *Random Forest* y *Support Vector Machines* tienen un menor nivel de créditos clasificados como *No Default* que en realidad sí hacen *default*.

Figura 9 Distribución de utilidad con una tasa de interés del 18% anual.

7. Conclusiones y recomendaciones.

Los resultados experimentales de este trabajo muestran que *Boosting* es la mejor estrategia para calibrar los modelos de clasificación de riesgo de crédito. Según el criterio del AUC, fue el mejor modelo de clasificación, seguido por el de Regresión Logística con penalización tipo *Lasso*. Los modelos de *Random Forest* y *SVM* son los de menor desempeño.

Se encontro que las variables más influyentes fueron las relacionadas a obligaciones adquiridas con el sector financiero.

Por otro lado, los resultados del criterio de la utilidad sitúan al modelo de *Random Forest* como el modelo que mayor utilidad le dejaría a las entidades que lo empleen. Esto se debe a que la incidencia del error de clasificación de créditos que hacen default es menor que en los otros modelos, lo que lleva a que el costo de prestar con este modelo sea menor.

Se tiene que los modelos *SVM y Random Forest* son modelos que se tienden a sobre ajustar y son muy sensibles a los cambios de individuos. Por lo tanto, al momento de emplearlos es fundamental usar técnicas de validación con datos por fuera de la muestra de entrenamiento.

Un aspecto que podría generar mejores resultados para estos modelos es tener un mayor contexto de la economía en la que se está trabajando y con esta información crear variables que permitan tener una mayor discriminación de los clientes riesgosos de los que no lo son. Adicional a esto contar con variables demográficas que permitan crear modelos para clientes que nunca han tenido experiencia crediticia.

8. Referencias.

Altman, Edward I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance: P. 189–209. V. 23. N 4

Anderson, Raymond. (2007) The Credit Scoring Toolkit. Oxford. P. 7.

Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang. (2007). Credit scoring with a data mining approach based on support vector machines, Expert Systems with Applications 33 P. 847–856.

Dobson, Annette. (2008) An Introduction to Generalized Linear Models. 3rd Edition. Chapman & Hall.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R. Springer.

Tomczak, Jakub, Maciej Zie¸ba. Classification Restricted Boltzmann Machine for Comprehensible Credit Scoring Model, Expert Systems with Applications 42 P.1789 – 1796.