CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

TALLER: TRABAJO FINAL INTEGRADOR

ESTIMACION DE LA PROBABILIDAD DE DEFAULT METODOLOGIAS ALTERNATIVAS

Nombre y Apellido del Alumno: Máximo Sangiácomo

Título de grado o posgrado (último): Magister en Finanzas

Profesores:

Alicia Mon

Juliana Gambini

Lugar y Fecha: CABA. 23 de Junio de 2020

1. Introducción

La irrupción de las firmas *BigTech* en la provisión de crédito está modificando la estructura del sistema financiero. Si bien la actividad principal de estas compañías es la provisión de servicios digitales como el e-commerce y servicios de pago paulatinamente han ido incorporando otros productos como la provisión de crédito, seguros, inversiones y ahorro.

El modelo de negocios de las *BigTech* difiere del modelo de las entidades financieras tradicionales principalmente por dos factores distintivos: efectos de red (generados por las plataformas de *e-commerce*, aplicaciones de mensajería y redes sociales); el uso de la tecnología (inteligencia artificial utilizando *big data*).

La utilización de nuevas técnicas de análisis y fuentes de datos alternativos brindan a las empresas tecnológicas una ventaja informativa para la evaluación de deudores respecto de las entidades financieras, que utilizan métodos econométricos convencionales (ej. estimaciones *logit*) menos flexibles para capturar la información contenida en grandes volúmenes de datos.

Mejorar la capacidad de evaluación de deudores para las entidades financieras es importante por varias razones. Primero, permite reducir el riesgo de crédito y con ello mejorar la solvencia de los bancos que a su vez impacta la estabilidad financiera del sistema. Segundo, en la medida que puedan incorporarse nuevas dimensiones de análisis permitirá la evaluación de sectores no atendidos aumentando la inclusión financiera. Tercero, permite emparejar oportunidades para los bancos.

Bajo la hipótesis que existen metodologías vinculadas al análisis de grandes volúmenes de datos que son sistemáticamente superiores, en términos de capacidad predictiva de la probabilidad de default de los clientes bancarios, respecto a los modelos *logit* generalmente utilizados por las entidades financieras, el objetivo del presente trabajo es determinar la metodología que brinde mejor performance predictiva comparando los resultados de *Gradient Boosting*, *Random Forest*, *Support Vector Machine* y la *Regresión Logística* tradicionalmente utilizada por los bancos.

Para ello, se cuenta con datos de la Central de Deudores del Banco Central de la República Argentina y del Ministerio de Producción de la Nación en distintos períodos tiempo. De esta forma, se busca brindar herramientas alternativas que permitan enriquecer la evaluación de clientes a las entidades financieras.

Los resultados indican que la *performance* predictiva de las metodologías de árboles de decisión es sustancialmente superior a la de la *Regresión Logística* y *Support Vector*

Machine dado que alcanzan valores de las métricas de análisis (área bajo la curva y sensibilidad) más elevados.

El documento se organiza como sigue. En la sección 2 se revisa la literatura. En la sección 3 se define el problema y en la sección 4 se justifica el estudio. En la sección 5 se detallan los alcances y limitaciones del trabajo. En las secciones 6 y 7 se presentan la hipótesis y los objetivos del estudio. La sección 8 describe cada una de las metodologías utilizadas y la sección 9 la base de datos, en la sección 10 se exhiben los resultados obtenidos. Finalmente, se presentan las conclusiones.

2. Revisión de la literatura

Tradicionalmente las entidades financieras han utilizado métodos econométricos convencionales (usualmente modelos *logit*) para estimar el riesgo de crédito. Los desarrollos tecnológicos de los últimos años han mejorado la capacidad de almacenar, procesar y analizar grandes volúmenes de datos. La literatura reciente se ha concentrado en estudiar técnicas de análisis que permitan explotar la riqueza contenida en la mayor disponibilidad de información mejorando la capacidad de predicción de los modelos de riesgo de crédito y su impacto sobre la inclusión financiera.

Respecto de la *performance* predictiva, Petropoulos y otros (2018) utilizan datos de préstamos a nivel de firmas del Banco Central de Grecia y aplican técnicas de árboles de decisión y redes neuronales comparadas a la estimación logística y análisis discriminante lineal para clasificar a los deudores y estimar su probabilidad de default. Los resultados señalan una mejor *performance* de las técnicas vinculadas a *machine learning* respecto de metodologías tradicionales.

En tanto, Khandani y otros (2010) aplican árboles de decisión y S*upport Vector Machines* a datos de préstamos al consumo de un banco cuyo nombre se mantiene anónimo donde combinan información tradicional (ej. la relación deuda / ingresos) con datos adicionales de transacciones (que normalmente no se encuentran disponibles) para obtener una medida del riesgo crediticio que mejora la capacidad predictiva respecto del *credit scoring* calculado por el banco.

Por su parte, de Castro Vieira y otros (2019) comparan la capacidad predictiva de modelos que estiman la probabilidad de default con métodos de *credit scoring* y técnicas *machine learning* sobre préstamos hipotecarios otorgados a familias de bajos ingresos en Brasil. Encuentran que los árboles de decisión basados *bagging* resultan los de mejor *performance* relativa y que aplicando dicha metodología para evaluar programas de financiamiento de vivienda podría reducirse la tasa de default del 11,8% a 2,95% lo que se traduce en una mejora de la calidad crediticia de la cartera y en reducción de pérdidas monetarias.

Otros estudios argumentan a favor de la riqueza de información que poseen las *BigTech* respecto de las entidades financieras. En tal sentido, Frost y otros (2019) muestran que el contenido informativo de los ratings internos de Mercado Crédito (perteneciente a Mercado Libre) basados en *big data* y *machine learning*, poseen una ventaja respecto de los *ratings* elaborados con la central de crédito tradicional. En particular, muestran que si la decisión de crédito hubiese estado basada solamente en datos de central de crédito el 30% de los clientes hubiera sido clasificado en la categoría de "riesgo alto" quedando excluidos del programa de crédito, mientras que ese valor es del 6% de acuerdo al *rating* interno.

A su vez, muestran que la calificación interna de Mercado Libre segmenta las originaciones en cinco grupos de riesgo diferentes (A a E) versus tres grupos identificados por la central de crédito. Esta mayor desagregación implica que, mientras para una calificación crediticia determinada por los datos de la central de crédito (ej. baja) la tasa de pérdida esperada es estrictamente monótona respecto de la calificación interna, no sucede lo mismo al fijar una calificación interna (ej. C, D o E). En este caso, la tasa de pérdida varía con el riesgo de la central de crédito, por ejemplo, existen casos donde para la calificación interna D la pérdida esperada en la categoría "riesgo bajo" de la central crédito es mayor que la pérdida esperada en la categoría "riesgo medio" permitiendo mayor granularidad en el análisis.

Por su parte, Gambacorta y otros (2019) utilizando datos de una *FinTech* china a nivel de empresas desde mayo a septiembre de 2017 comparan el poder predictivo para la probabilidad de default (con metodología *logit*) de tres modelos alternativos. El primero utiliza solamente el resultado del score de crédito obtenido por la *FinTech* con métodos de *machine learning*. El segundo incorpora únicamente información tradicional que normalmente está disponible para los bancos. El tercero le suma a este último variables no tradicionales obtenidas por la *FinTech* a través de aplicaciones móviles y la plataforma de *e-commerce*. Los resultados indican que el primer modelo tiene la mejor *performance* (mayor AUROC), seguido por el tercer modelo quedando el segundo modelo en último lugar, señalando nuevamente una ventaja informativa de la *FinTech* en el procesamiento de información respecto de los bancos.

La capacidad de analizar información no tradicional permite ampliar el universo de potenciales clientes. En general, las empresas PyMEs tienen dificultades para cumplir con los requisitos de información formal (ej. datos de balance, garantías) exigidas por las entidades financieras, quedando excluidas de los mercados de crédito. A su vez, esto produce un círculo vicioso porque les impide crear una historia crediticia que señale su calidad como deudor. Las empresas *BigTech* pueden explotar información de plataformas de *e-commerce*, aplicaciones móviles y redes sociales para extraer datos vinculados a volúmenes de venta, participación de mercado, valoración por parte de los clientes permitiendo analizar y evaluar las características de sus clientes.

En este sentido, Bazarbash M. (2019) analiza las fortalezas y debilidades de la evaluación de créditos con técnicas de *machine learning*. Señala que las mismas pueden contribuir a incrementar la inclusión financiera ya que, por un lado, permitirían reducir los costos y, por el otro, aprovechar el acceso a información no tradicional para mejorar la evaluación de garantías y la proyección de ingresos.

Por su parte, Jagtiani y Lemieux (2017 y 2018) comparan datos de préstamos realizados por la *FinTech Lending Club* y datos de tarjetas de crédito de los reportes Y-14M de la Reserva Federal de Estados Unidos. Encuentran que los préstamos al consumo de *Lending Club* penetran en áreas que se pueden beneficiar de la mayor oferta de crédito (ej. zonas donde se redujo la cantidad de sucursales bancarias). Asimismo, el uso de información no tradicional permite mejorar la clasificación y reducir los *spreads* de algunos deudores respecto de los criterios tradicionales.

En términos de eficiencia, Fuster y otros (2018) encuentran que la innovación tecnológica de las *FinTech* ha mejorado la intermediación financiera en el mercado hipotecario Estados Unidos. Sus resultados indican que estas compañías procesan solicitudes aproximadamente un 20% más rápido que otros prestamistas. Además, son capaces de ajustar la oferta de manera más elástica en respuesta a shocks de demanda, aliviando así las limitaciones de capacidad asociadas con los préstamos hipotecarios tradicionales.

3. Definición del problema

La irrupción de las firmas *BigTech* en la provisión de crédito está modificando la estructura del sistema financiero. Las mismas se definen como empresas cuya actividad principal es la brindar servicios digitales, principalmente *e-commerce* y servicios de pago, que luego han ido incorporando otros productos vinculados al otorgamiento de préstamos, seguros, inversiones y ahorro.

El modelo de negocios de las *BigTech* difiere del modelo de las entidades financieras tradicionales principalmente por dos factores distintivos: efectos de red (generados por las plataformas de *e-commerce*, aplicaciones de mensajería y redes sociales); el uso de la tecnología (inteligencia artificial utilizando *big data*).

La utilización de nuevas técnicas de análisis y fuentes de datos alternativos brindan a las empresas tecnológicas una ventaja informativa para la evaluación de deudores respecto de las entidades financieras, dado que usualmente basan sus decisiones en métodos econométricos convencionales (ej. estimaciones *logit*) que resultan menos flexibles para poder capturar de manera eficiente la información contenida en grandes volúmenes de datos.

4. Justificación del estudio

Mejorar la capacidad de evaluación de deudores para las entidades financieras es importante por varias razones. Primero, permite reducir el riesgo de crédito y con ello mejorar la solvencia de los bancos que a su vez impacta la estabilidad financiera del sistema. Segundo, en la medida que puedan incorporarse nuevas dimensiones de análisis permitirá la evaluación de sectores no atendidos aumentando la inclusión financiera. Tercero, permite emparejar oportunidades para los bancos.

5. Alcances y limitaciones

La base de datos utiliza información mensual confidencial a nivel de firmas para los períodos Marzo a Diciembre de 2017 y Enero a Octubre de 2019 e integra dos fuentes alternativas. Por un lado, la Central de Deudores del Banco Central de la República Argentina registra información sobre montos y características de deuda en el sistema bancario. Por el otro, datos de montos exportados y países de destino (información no tradicional para los bancos) provenientes del Ministerio de Producción de la Nación.

Es importante señalar como limitación del trabajo que no dispone de información de balance para calcular indicadores tradicionales que permitirían una caracterización más detallada de las firmas.

6. Hipótesis

Existen metodologías vinculadas al análisis de grandes volúmenes de datos que sistemáticamente son superiores en términos de capacidad predictiva de la probabilidad de default de los clientes respecto a los modelos *logit* generalmente utilizados por las entidades financieras.

7. Objetivo general

Determinar la metodología para estimar la probabilidad de default de los clientes bancarios que brinde mejor performance.

Objetivos específicos:

1. Seleccionar y consolidar las bases de datos a nivel de firmas provenientes de la Central de Deudores del Banco Central de la República Argentina y del Ministerio de Producción de la Nación.

2. Aplicar la metodología de ***Gradient Boosting*** y evaluar la performance predictiva de la probabilidad de default en la muestra seleccionada para testing.
3. Aplicar la metodología de ***Random Forest*** y evaluar la performance predictiva de la probabilidad de default en la muestra seleccionada para testing.
4. Aplicar la metodología de ***Support Vector Machine*** y evaluar la performance predictiva de la probabilidad de default en la muestra seleccionada para testing.
5. Aplicar la metodología de **Regresión Logística** y evaluar la performance predictiva de la probabilidad de default en la muestra seleccionada para testing.
6. Comparar resultados de las predicciones anteriores a través de distintas métricas para establecer un ranking y así determinar la metodología con mejor performance predictiva de la probabilidad de default de clientes bancarios.

8. Metodología

En lo que sigue se describen brevemente las características principales de las metodologías de estimación utilizadas sin ahondar en los detalles técnicos dado que exceden al objetivo de este trabajo.

Los modelos *logit* resultan apropiados cuando la variable dependiente es cualitativa. Por ejemplo, al estimar la probabilidad de default de una empresa donde la categoría toma solo dos valores, SI o NO, se puede adoptar el enfoque de las variables binarias (0/1) reemplazando el Si por un 1 y al No por un 0. En dicho caso, la metodología *logit* estima la probabilidad de ocurrencia de un evento condicional en un conjunto de información.

Formalmente, se busca estimar $p(X) = Pr(Y = 1/X)$. Si bien podría utilizarse un modelo lineal:

$$p(X) = \beta_0 + \beta_1 X \qquad (1)$$

esta especificación presenta un serio problema. El rango de la variable estimada no está acotado al intervalo [0, 1] pudiendo resultar en valores negativos o mayores a 1 contraintuitivos para una probabilidad.

Para evitar este problema, la regresión logística utiliza la función logística:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad (2)$$

cuyos parámetros se estiman por el método de máxima verosimilitud. En dicha especificación, los valores de respuesta quedan acotados al intervalo [0, 1] y luego pueden

mapearse a un modelo de clasificación utilizando una regla del tipo del clasificador de Bayes. Por ejemplo, si p(Y) > 0,5 el caso pertenece a la clase de default.[1]

Por otro lado, los árboles de decisión están compuestos por una raíz, nodos intermedios y nodos terminales. El algoritmo divide el espacio de variables independientes de manera jerárquica y en forma secuencial desde la raíz hasta llegar a los nodos terminales, de esta manera, dentro de cada rama las últimas decisiones dependen de las primeras y las observaciones se dividen en regiones no superpuestas. Luego, para cada observación que cae en una determinada región se realiza la misma predicción basado en la media de la variable dependiente (o la clase con mayor proporción de observaciones en el caso de clasificación) en los datos de entrenamiento.

El objetivo es encontrar regiones de manera de minimizar la suma de residuos al cuadrado o la tasa de error de clasificación definida como la fracción de las observaciones de la muestra de entrenamiento que no pertenecen a la clase mayoritaria de una determinada región. En el caso de regresión se busca minimizar:

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_{R_j})^2 \qquad (3)$$

donde y_i es la respuesta observada y $\hat{y}_{R_j}$ es la respuesta media en el nodo terminal R_j en la muestra de entrenamiento y J la cantidad total de nodos.

La construcción de un solo árbol puede llevar a soluciones muy complejas tendientes a hacer *overfitting*, es decir, identificar patrones en la muestra de entrenamiento que en ocasiones pueden ser causados por cuestiones aleatorias en vez de ser patrones verdaderos. En dichos casos, los errores en la muestra de testeo resultan elevados porque los patrones encontrados en entrenamiento no aparecen en dicha muestra, es decir el modelo generaliza poco.

Por ello, la metodología *Random Forest* busca solucionar el problema anterior en base a al principio de *bagging* con árboles no correlacionados. El *bagging* estima varios árboles en submuestras seleccionadas con reposición de los datos de entrenamiento (también conocido como *bootstrapping*) donde la predicción final es el promedio de las estimaciones o el voto de la mayoría en el caso de clasificación. Para disminuir la correlación de los árboles estimados, en cada partición sólo se considera, en forma aleatoria, un subconjunto de variables independientes evitando el problema de variables dominantes que generan árboles similares.

Por su parte, *Gradient Boosting* utiliza una estrategia de estimación alternativa ya que en vez de construir y combinar árboles paralelos e independientes como *Random Forest*

[1] El umbral seleccionado puede variar de acuerdo al problema que se quiera estudiar.

construye una serie de árboles, donde cada uno “aprende” en base a los errores del árbol anterior. Primero, se estima un árbol simple con la muestra de entrenamiento. Luego, se estiman árboles secuencialmente utilizando los errores de predicción del árbol previo como la variable independiente para ir reduciendo el error de predicción. La predicción final es la suma ponderada de las predicciones de todos los árboles. La ponderación se rige por un parámetro que regula la velocidad de aprendizaje.

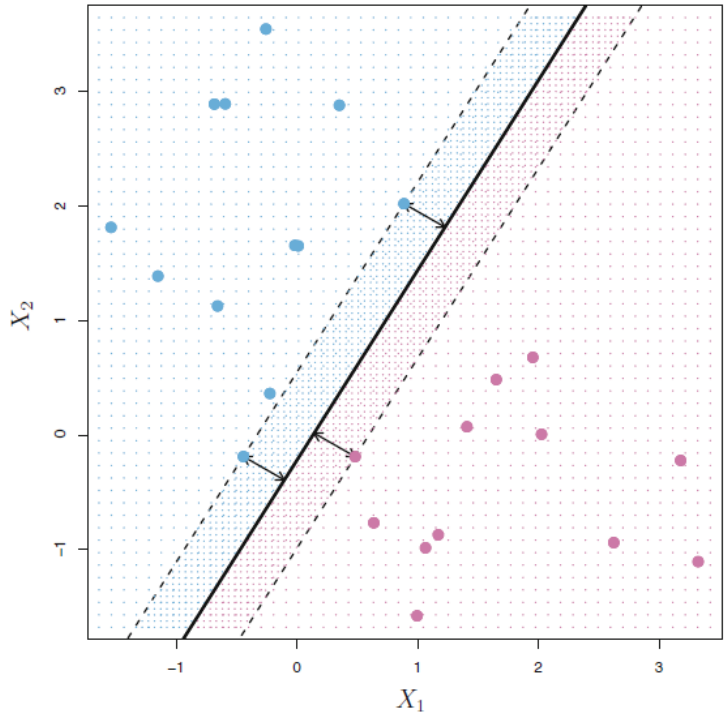
De acuerdo a James y otros (2017), *Support Vector Machine* (SVM) es una generalización de un clasificador llamado “margen máximo” basado en una idea sencilla. Supone que dado un conjunto de observaciones representadas en *p* dimensiones pertenecientes a dos clases diferentes existe un hiperplano de dimensión *p-1* que puede ubicarlas en regiones separadas. Claramente en la práctica la separación no se puede realizar de manera perfecta así que el algoritmo contempla estrategias donde sacrifica la clasificación de algunas observaciones en búsqueda de mejor ajuste posible.

Este método busca maximizar la distancia perpendicular mínima entre el hiperplano de separación y los puntos de cada clase formando bandas conocidas como margen. Se espera que un clasificador que tenga un margen máximo amplio en los datos de entrenamiento también tendrá un margen amplio en los datos de test y, por lo tanto, clasificará correctamente las observaciones en la muestra de testeo.

La Figura 1 presenta un ejemplo gráfico de S*upport Vector Machine.* Exhibe la variable *target* que posee dos clases azul y violeta en función de 2 variables predictoras X_1 y X_2. La línea gruesa continua representa el hiperplano de margen máximo (en el caso de 2 dimensiones el hiperplano es una recta). El margen es la distancia desde esta última hasta las líneas punteadas señalado con flechas. Los dos puntos que se encuentran en azul y violeta ubicados sobre líneas discontinuas son los vectores de soporte (porque en un plano p-dimensional esos puntos son vectores) dado que soportan al margen máximo.[2] Las áreas de colores indican la regla de decisión de la SVM. Finalmente, existen distintas formas funcionales para el hiperplano de separación conocidas como *kernels* que brindan mayor flexibilidad al algoritmo.

[2] Notar que si los puntos se mueven levemente la línea continua también se moverá por lo que el hiperplano depende de pocas observaciones.

Figura 1. Ejemplo de aplicación de *SVM* a un conjunto de datos con dos clases.



Fuente: James y otros (2017)

8.1. Estrategia de estimación

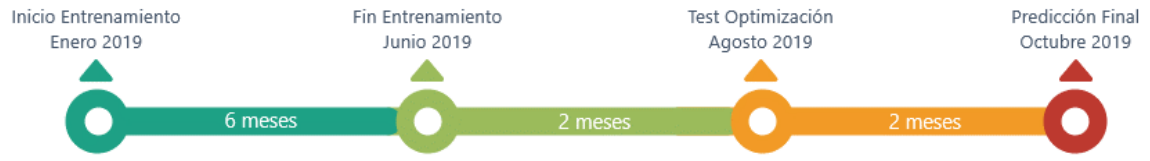
El ejercicio busca estimar la probabilidad de que una firma entre en situación de default en los próximos 2 meses. Para analizar la calidad de las predicciones suele dividirse la muestra en dos partes, una de ellas es utilizada para entrenar el modelo y la otra para hacer el testeo de las predicciones. Esto es así porque al elaborar un modelo de predicción el interés está puesto en cómo va a funcionar en datos nuevos, es decir, que no hayan sido "vistos" previamente por el modelo.

Por otro lado, dado que los resultados de los árboles de decisión dependen de la configuración de los hiperparámetros de cada algoritmo, existen distintas metodologías para buscar los valores más adecuados. En el presente trabajo se utiliza un método de optimización bayesiana donde se busca el modelo que maximice el área bajo la curva (AUC por sus siglas en inglés) en la muestra de testeo.

Para ello, la estrategia de estimación consiste primero en fijar el mes a utilizar para realizar las predicciones, Diciembre para el año 2017 y Octubre el año 2019. Luego, parado en ese momento del tiempo se deja un espacio de 2 meses hacia atrás y se toma esa fecha como

la muestra de testeo del ejercicio de optimización y, dado que se busca estimar la probabilidad de default en los próximos 2 meses, se deja nuevamente un período de 2 meses hacia atrás. A partir de esa fecha se establecen 6 meses consecutivos donde es posible realizar el entrenamiento (Figura 2). De esta forma, el algoritmo de optimización bayesiana se encarga de definir el valor de los hiperparámetros y la cantidad de meses óptimo para entrenar en 35 iteraciones buscando maximizar el AUC.

Figura 2. Línea temporal del proceso de optimización.



En general, al realizar predicciones, un clasificador va a tener aciertos y errores. A su vez, tales resultados dependen del umbral establecido para asignar la probabilidad estimada a una categoría específica. La curva ROC muestra la tasa de verdaderos positivos (TPR, por sus siglas en inglés) contra la tasa falsos positivos (FPR) para distintos valores de ese umbral. El TPR también se conoce sensibilidad y la FPR como (1 - especificidad). El área bajo la curva ROC (AUC) varía del 50% predicción puramente aleatoria, al 100% predicción perfecta.

En tanto, para la metodología de *Support Vector Machine* se probaron distintos valores del parámetro de costo y distintas especificaciones de kernel[3] seleccionando la especificación con mayor AUC (kernel sigmoide y el costo igual 1) para realzar el ejercicio de predicciones. Finalmente, la Regresión Logística no posee parámetros para ajustar. En ambos casos se utilizaron 6 meses para entrenar.

Si bien el AUC es el principal criterio de decisión del presente trabajo, más adelante se mostrarán otras 3 medidas básicas que contribuyen a evaluar la performance del modelo:

$$Precision = \frac{VP+VN}{VP+VN+FP+FN} \qquad (4)$$

$$Sensibilidad = \frac{VP}{VP+FN} \qquad (5)$$

[3] Sigmoide, lineal y polinomio de grado 3.

$$Especificidad = \frac{VN}{VN+FP} \quad (6)$$

donde *VP* representa a los verdaderos positivos, *VN* verdaderos negativos, *FP* los falsos positivos y *FN* los falsos negativos en una matriz de confusión.

La matriz de confusión (Tabla 1) compara las predicciones del modelo con los datos observados en la base de *test*. Por lo tanto, los elementos de la diagonal principal muestran los casos cuyos valores fueron correctamente predichos (la *Precisión* o *accuracy* es la suma de estos elementos sobre el total de casos).

La *Sensibilidad* indica la cantidad de aciertos sobre los casos posibles dentro de cada categoría, mientras que la *Especificidad* señala el complemento de las observaciones que, no perteneciendo a una clase, fueron clasificadas dentro de la misma. Dicho de otra manera, es el grado de confianza que puedo depositar en que la clasificación dentro de una clase sea correcta. En el presente caso de estudio, la sensibilidad muestra el porcentaje de empresas que hicieron default correctamente identificadas y la especificidad es la proporción de empresas que no incumplieron sus pagos correctamente identificadas.

Tabla 1. Matriz de confusión.

		Predicho		**Total**
		NO	**SI**	
Observado	**NO**	VN	FP	**N**
	SI	FN	VP	**P**
	Total	**N***	**P***	

En resumen, dado que uno de los meses que se utilizó para hacer las predicciones es Octubre de 2019, se retrocede hasta Agosto de 2019 y ese mes es el de testeo para el proceso de optimización bayesiana. Luego se retroceden 2 meses más hasta Junio de 2019 y a partir de allí se toman un período de 6 meses consecutivos hasta Enero de 2019 como muestra, donde se buscan los hiperparámetros y la cantidad de meses de entrenamiento. Finalmente, con los valores definidos se estima la probabilidad de default en el mes de establecido para el análisis. La Tabla 2 presenta el rango valores de los hiperparámetros a optimizar en cada algoritmo.

Tabla 2. Rango de hiperparámetros optimizados en cada metodología.

Metodología	Parámetro	Definición parámetro	Rango	Tipo
Gradient Boosting	max.depth	Profundidad máxima del árbol	2 - 50	Entero
	eta	Parámetro que controla la tasa de aprendizaje	0.001 - 0.99	Numérico
	nrounds	Número máximo de iteraciones	2 - 50	Entero
	cantidad de meses		1 - 6	Entero
Random Forest	num.trees	Número de árboles	25 - 200	Entero
	min.node.size	Tamaño mínimo del nodo	1 - 100	Entero
	mtry	Número de variables posibles para dividir cada nodo	2 - 7	Entero
	cantidad de meses		1 - 6	Entero

9. Base de datos

La base de datos utiliza información mensual confidencial a nivel de firmas para los períodos Marzo a Diciembre de 2017 y Enero a Octubre de 2019 e integra dos fuentes alternativas (en total se dispone de 215.726 empresas diferentes para el primer año y 237.090 para el segundo). Por un lado, la Central de Deudores del Banco Central de la República Argentina registra información sobre montos y características de deuda en el sistema bancario. Por el otro, datos de montos exportados y países de destino (información no tradicional para los bancos) provenientes del Ministerio de Producción de la Nación.

Es importante señalar como limitación del trabajo que no dispone de información de balance para calcular indicadores tradicionales que permitirían una caracterización más detallada de las firmas.

La Central de Deudores cuenta con datos mensuales de endeudamiento a nivel de firmas (el presente estudio se concentra en el sector privado no financiero). Además del saldo de la asistencia recibida (teniendo en cuenta sólo crédito efectivamente tomado por la empresa y no el total de financiaciones otorgadas, por lo que se excluye la tenencia de obligaciones negociables, acciones, la extensión de garantías y las líneas de préstamos puestas a disposición pero no utilizadas por la firma) el régimen informativo identifica a: i) la clasificación del deudor (situación 1 a 6); ii) la entidad otorgante del préstamo; iii) la línea de asistencia, iv) el tipo de garantía que respalda la operación, v) la apertura por moneda

de dicha asistencia; vi) la actividad principal del deudor y; vii) si existe una relación de vínculo entre el acreedor y el deudor.

De esta forma, para estimar la probabilidad de default la variable dependiente (o *target*) es binaria, toma el valor 1 en caso que en un mes determinado la situación máxima asignada por alguna de las entidades que asisten a la firma tome el valor en el rango 3 a 6 y 0 en caso contrario.

Por el lado del endeudamiento, en base a Bebczuk y Sangiácomo (2008), se construyeron las siguientes variables independientes: el logaritmo del saldo de deuda mensual en miles de pesos, el porcentaje de deuda que cuenta con garantías, la proporción de deuda en moneda extranjera y la concentración de deuda medida por el Índice Herfindahl-Hirschman (HHI).[4]

Respecto de las entidades financieras que asisten al deudor se calculó el número total de entidades por mes, se identificó la entidad que otorga la máxima asistencia y el grupo económico al que pertenece dicha entidad de acuerdo a su estructura de capital: Bancos Públicos, Bancos Locales de Capital Nacional, Bancos Locales de Capital Extranjero y Bancos Sucursales de Entidades Financieras del Exterior.

Por el lado del tipo de asistencia, se identifica el total de líneas diferentes que recibe el deudor y la línea de máxima asistencia en un determinado mes.

En relación a las características del deudor, se dispone del sector de actividad en donde opera,[5] una variable *dummy* que registra si la firma es exportadora (o no), la cantidad de países de destino y una variable binaria que toma el valor 1 si en un mes determinado más del 40% de las exportaciones se destinan a mercados desarrollados.

Las Tablas 3 y 4 presentan la cantidad de firmas por mes en cada una de las muestras utilizadas clasificadas según la situación de default. Se advierte que la tasa promedio es del 5,3% en 2017 y algo más elevada, 7,4% en 2019.

[4] Para cada mes se calcula la suma de la proporción de deuda otorgada por las entidades que asisten a la firma -m*arket-share*- elevada al cuadrado, un valor igual a 1 implica total concentración de deuda en una sola entidad y a medida que el HHI se acerca a 0 la concentración disminuye.

[5] Se eliminaron las firmas pertenecientes al sector financiero y al Sector Público dado que su operatoria difiere de las firmas productoras de bienes y servicios.

Tabla 3. Cantidad de firmas por situación año 2017.

Mes	No default	Default	Total	% de default
2017m3	156.054	7.579	163.633	4,63
2017m4	156.964	7.556	164.520	4,59
2017m5	156.157	8.448	164.605	5,13
2017m6	157.025	8.208	165.233	4,97
2017m7	134.399	8.241	142.640	5,78
2017m8	134.098	7.995	142.093	5,63
2017m9	135.284	7.796	143.080	5,45
2017m10	136.138	8.142	144.280	5,64
2017m12	137.921	8.596	146.517	5,87

Tabla 4. Cantidad de firmas por situación año 2019.

Mes	No default	Default	Total	% de default
2019m1	160.790	12.555	173.345	7,24
2019m2	161.978	12.370	174.348	7,10
2019m3	163.747	12.469	176.216	7,08
2019m4	164.242	12.324	176.566	6,98
2019m5	163.911	13.297	177.208	7,50
2019m6	164.048	13.547	177.595	7,63
2019m7	162.891	13.459	176.350	7,63
2019m8	164.170	13.288	177.458	7,49
2019m10	164.471	13.604	178.075	7,64

10. Resultados de las estimaciones

Las Figuras 3 y 4 muestran la evolución de la métrica AUC en *testing* del procedimiento descrito en la sección precedente para los distinto períodos y metodologías, a lo largo de las iteraciones (rojo) y el valor máximo alcanzado (verde) en relación al tiempo de procesamiento medido en horas. A su vez, registran el total de iteraciones y el valor de AUC máxima.

Se observa que si bien los valores de esta última métrica alcanzados por cada metodología en cada mes utilizado para test no difieren demasiado (0.93 para 2017m10 y 0.94 para 2019m8), existe una diferencia importante en los tiempos de estimación. Mientras que *Gradient Boosting* demoró alrededor de 3 horas en realizar 35 iteraciones *Random Forest* lo hizo en 20 horas para 2017m10 y 9 horas para 2019m8, por lo que la primera es significativamente más eficiente que la segunda. Además, se observa que a partir de un determinado número de iteraciones los algoritmos “aprenden” y los valores de AUC se mantienen relativamente estables.

Figura 3. Evolución de AUC en la optimización para 2017m10

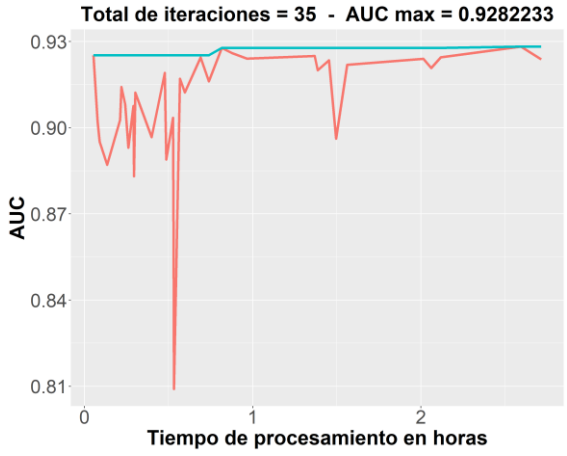
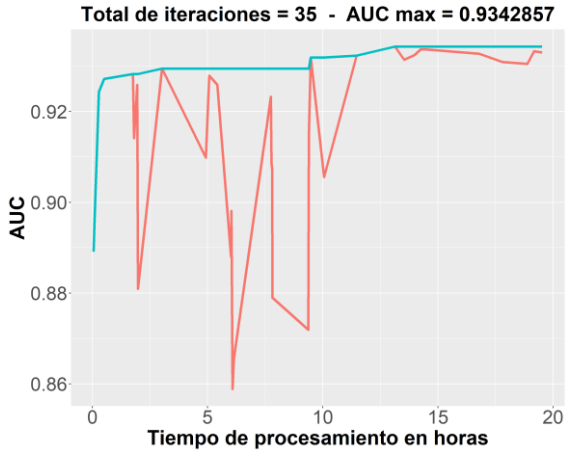
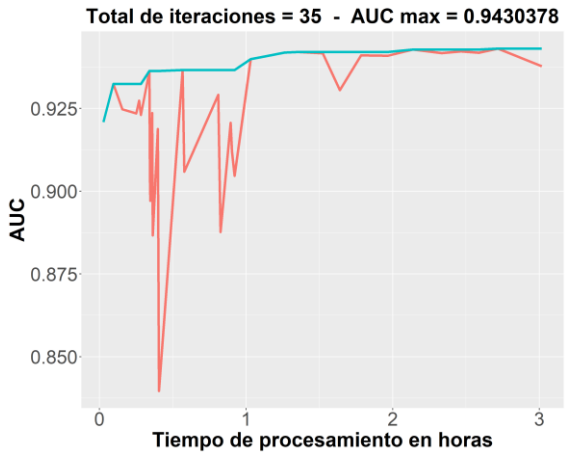
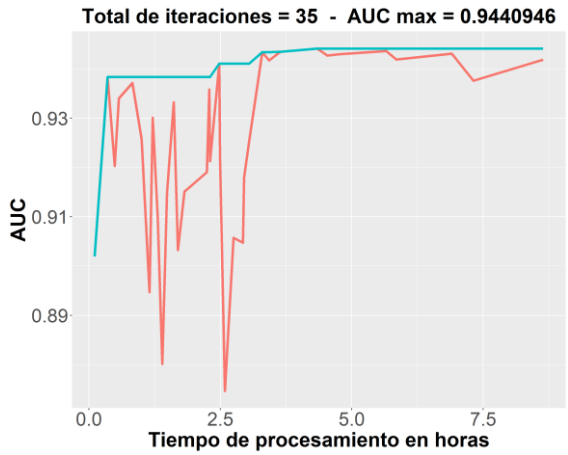
Gradient Boosting	Random Forest
	

Figura 4. Evolución de AUC en la optimización para 2019m08

Gradient Boosting	Random Forest
	

La Tabla 5 exhibe los valores de los hiperparámentros resultantes del proceso de optimización que se utilizaron para realizar la predicción final para cada mes correspondiente.

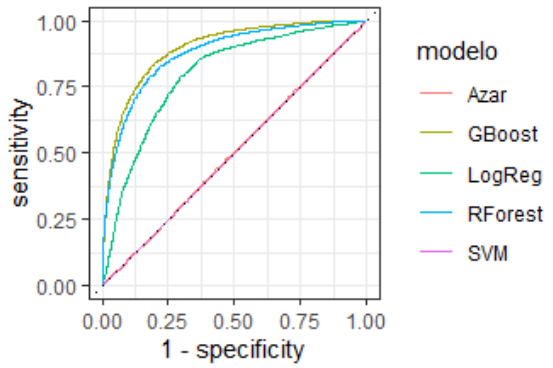
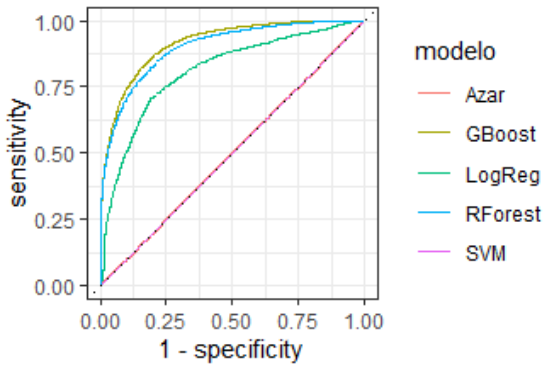
Tabla 5. Valores de hiperparámetros optimizados en cada metodología.

Metodología	Parámetro	2017m12	2019m10
Gradient Boosting	max.depth	47	42
	eta	0.21223	0.14416
	nrounds	38	49
	cantidad de meses	6	6
Random Forest	num.trees	195	187
	min.node.size	7	9
	mtry	5	6
	cantidad de meses	4	5

La Figura 5 presenta la curva AUC para las distintas metodologías y períodos de estimación. Los resultados indican un patrón bastante claro. Primero, las metodologías vinculadas a los árboles de decisión son las que obtuvieron la mejor *performance* (las curvas de *Gradient Boosting* y *Random Forest* prácticamente están superpuestas). En el otro extremo quedó SVM que no logra diferenciarse de un modelo completamente aleatorio. Finalmente, la Regresión Logística se ubica entre estos dos casos extremos quedando en tercer lugar de consideración. Por lo tanto, de las 3 metodologías vinculadas a *machine learning* las dos

primeras lograron superar a la Regresión Logística en la estimación de la probabilidad de default.

Figura 5. Evolución de AUC para cada metodología y período

2017m12	2019m10
	

Por su parte, las tablas 6 y 7 exhiben distintas métricas que también sirven para analizar los resultados de las estimaciones para 2017m12 y 2019m10, respectivamente. Además del área bajo la curva que constituye el criterio de decisión principal de este trabajo, se muestra la precisión, sensibilidad y especificidad.

De las 3 medidas adicionales, en el análisis de la probabilidad de default de empresas la sensibilidad resulta crucial dado que para un banco es muy importante identificar como default a las empresas que realmente van a incumplir con sus pagos mientras que clasificar de manera incorrecta a una firma cumplidora reviste menor gravedad.

Por otra parte, si bien la precisión es una medida usualmente utilizada para medir la *performance* de predicción los modelos, en el presente caso de estudio donde las tasas de default observadas son bajas (en promedio 5,3% en 2017 y 7,4% en 2019) un clasificador trivial que predice todos casos como no default (categoría igual a 0) va a tener una precisión elevada, siendo su error equivalente al complemento de la tasa de default observada en cada muestra. Por lo tanto, en los casos donde el valor de AUC resulta similar, se tomó a la sensibilidad como el segundo criterio de decisión para medir la *performance* de las predicciones.

Ambas tablas refuerzan los resultados anteriores, aunque permiten hacer una diferenciación entre la metodología de *Gradient Boosting* y *Random Forest*. Como era de esperar no existen diferencias significativas en los valores de AUC alcanzados en cada caso aunque, tanto para 2017m12 como para 2019m10, la primera metodología posee mayor sensibilidad que la segunda pudiendo establecer un ranking en los resultados finales

entre las distintas metodologías.[6] Primero se ubica *Gradient Boosting*, segundo *Random Forest*, la *Regresión Logística* queda en tercer lugar y *Support Vector Machine* no logra diferenciar los casos de default.[7]

En este sentido, la Regresión Logística y *Support Vector Machine* tienen una *performance* pobre dado que no se diferenciaron de un clasificador trivial ubicando a todas las firmas en la clase de no default resultando en una sensibilidad igual a 0 y una especificidad de 1 y la precisión es el complemento de la tasa de default observada en los meses respectivos (véase tablas 1 y 2).

Tabla 6. Métricas por metodología para 2017m12

Modelo	AUC	Precisión	Sensibilidad	Especificidad
Gradient Boosting	**0,896**	0,947	**0,241**	0,991
Random Forest	**0,877**	0,942	0,023	1
Regresión Logística	0,797	0,941	0,005	1
Support Vector Machine	0,500	0,941	0	1

Tabla 7. Métricas por metodología para 2019m10

Modelo	AUC	Precisión	Sensibilidad	Especificidad
Gradient Boosting	**0,915**	0,940	**0,389**	0,985
Random Forest	**0,900**	0,929	0,083	0,999
Regresión Logística	0,809	0,924	0,017	0,999
Support Vector Machine	0,500	0,925	0	1

Finalmente, los resultados alcanzados en el presente estudio están en línea con los obtenidos por Petropoulos y otros (2018) donde *Gradient Boosting* posee la mejor performance relativa, seguida por una red neuronal profunda (librería MxNET en R) quedando la Regresión Logística y el Análisis Discriminante Lineal (utilizados como *benchmark* de metodologías tradicionales) rezagados al último lugar.

[6] La sensibilidad alcanzada en general resulta baja. Probablemente esté vinculado a la limitación de información de no disponer de datos de balance de las firmas.
[7] A estas métricas se suma la eficiencia de *Gradient Boosting* para realizar las estimaciones, que si bien no se vincula a las medidas de *performance* de estimaciones, constituye característica deseable al momento de trabajar con grandes volúmenes de datos.

11. Conclusiones

El objetivo del presente trabajo fue determinar cuál es la metodología de estimación más apropiada para predecir la probabilidad de default de los clientes bancarios. Para ello, se compara la performance predictiva de cuatro metodologías de clasificación alternativas, tres de ellas relacionadas al análisis de grandes volúmenes de datos: *Random Forest*, *Gradient Boosting*, *Support Vector Machine* y, la cuarta, la Regresión Logística tradicionalmente utilizada por los bancos.

Para ello, se utilizaron datos de la Central de Deudores del Banco Central de la República Argentina en conjunto con información del Ministerio de Producción de la Nación en distintos períodos de tiempo.

Se ha logrado establecer un ranking en la *performance* predictiva de cada metodología basado los resultados alcanzados por las métricas de área bajo la curva y sensibilidad. De esta forma *Gradient Boosting* quedó ubicado en primer lugar, *Random Forest* en segundo puesto, la Regresión Logística tercero y finalmente *Support Vector Machine*.

Por lo tanto, de las 3 metodologías vinculadas a *machine learning* las dos primeras lograron superar a la Regresión Logística en la estimación de la probabilidad de default. A la luz de estos resultados, hacia adelante será importante para los bancos tradicionales explorar las técnicas de árboles de decisión como herramientas alternativas que permitan enriquecer la evaluación de clientes, reducir el riesgo de crédito y aumentar la inclusión financiera.

Bibliografía

Bazarbash M. (2019). “FinTech in Financial Inclusion Machine Learning Applications in Assessing Credit Risk.” IMF Working Paper 19/109.

Bebczuk R. y M. Sangiácomo (2008). “Determinantes de la cartera irregular de los bancos en Argentina.” Ensayos Económicos 51. Banco Central de la República Argentina.

de Castro Vieira J. R., F. Barboza, V. A. Sobreiro, H. Kimura (2019). “Machine learning models for credit analysis improvements: Predicting low-income families’ default.” Applied Soft Computing. Volume 83.

Frost J., L. Gambacorta, Y. Huang, H. Song Shin and P. Zbinden (2019). “BigTech and the changing structure of financial intermediation.” BIS Working Papers No 779.

Fuster A., M. Plosser, P. Schnabl, and J. Vickery (2018). “The role of technology in mortgage lending.” Working Paper 24500, National Bureau of Economic Research.

Gambacorta L., Y. Huang, H. Qiu† and J. Wang (2019). “How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm.” BIS Working Papers No 834.

Jagtiani J. and C. Lemieux (2017). “Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information.”

Jagtiani J. and C. Lemieux (2018). “Do fintech lenders penetrate areas that are underserved by traditional banks.” Federal Reserve Bank of Philadelphia Working Papers.

James G., D. Witten, T. Hastie and R. Tibshirani (2017). An Introduction to Statistical Learning with Applications in R. Springer.

Khandani A. E., A. J. Kim, and A. W. Lo (2010). “Consumer credit-risk models via machine-learning algorithms.” Journal of Banking & Finance 34 (2010): 2767-2787.

Petropoulos A., V. Siakoulis, E. Stavroulakis and A. Klamargias (2018). “A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting.” Bank of Greece.