DETERMINANTES DEL RIESGO DE INCUMPLIMIENTO EN CRÉDITOS EDUCATIVOS: UN ANÁLISIS PARA COLOMBIA

Manuela Granda

RESUMEN

En este documento se utilizan metodologías no paramétricas de *Machine Learning*, en particular el algoritmo XGBoost, para predecir el riesgo de incumplimiento en los créditos educativos en Colombia ofrecidos por el ICETEX entre los años 2015 a 2018. La variable de interes es el riesgo de incumplimiento en créditos estudiantiles (default) y se utilizan como determinantes variables asociadas al nivel socioeconómico de los estudiantes, así como la información del colegio y el logro académico de cada estudiante. Los principlaes resultados muestran que las variables socieconomicas con muy buenos predictores del default, en partiucular, variables como la educación de los padres y los puntajes en las pruebas de lectura crítica son fuertes predictores. Los resultados encontrados contribuyen a las decisiones de política económica y social sobre el diseño de métodos para la cobertura de educación superior a través de créditos meritorios con fondos públicos y privados.

JEL: I22, I21, D86

Palabras clave: créditos educativos, XGBoost, educación, educación superior.

I. INTRODUCCIÓN

La educación superior constituye uno de los principales factores productivos para la formación de capital humano, necesaria para el desarrollo económico y social de un país, de manera que los países con mejor educación superior han evidenciado tener mejores comportamientos del producto interno bruto (Barro, 1995).

La demanda por educación superior ha tenido una expansión significativa en todo el mundo, especialmente en países en desarrollo (Ciro, 2017). Sin embargo, la oferta de cupos nuevos en las universidades no ha tenido la misma velocidad de expansión (Ferreyra, 2017; Gómez y Celis, 2009).

Es así como las barreras existentes en el proceso de acceso a educación superior se convierten entonces en uno de los temas principales sobre la agenda política y social con el fin de incrementar cobertura y calidad en el proceso de formación profesional de los individuos. Existen diferentes métodos de acceso a educación superior alrededor del mundo. En países desarrollados el mercado de créditos educativos está altamente desarrollado y extiende una gama de numerosas alternativas para financiar diferentes programas profesionales, además la inversión pública de cada gobierno ha dado lugar a diferentes opciones de acceso, a través de beneficios parciales o totales sobre los costos financieros, a aquellas poblaciones vulnerables en términos de calidad de vida.

En países de ingreso medio y bajo los mercados de crédito educativo tienen desarrollo parcial. Específicamente, en América Latina se han adelantado los esfuerzos en materia de la asignación de recursos para este sector, lo cual ha llevado a que la cobertura en educación superior sufriera una gran expansión en términos de la demanda. Según cifras del Banco Mundial, el porcentaje de individuos entre los 18 y 24 años inscritos en educación superior en América Latina creció de 21 % en 2000 a 40 % en 2010. Dentro de los países que más han aportado a estas cifras se encuentra Colombia. La tasa de cobertura (porcentaje de individuos entre los 15 y los 24 años inscritos en educación superior) pasó de 25 % a 54 % en 2018. Sin embargo, los aportes públicos a la educación superior son tema de constante debate económico dado que en algunos casos resulta ser una medida regresiva (Ruiz, 1997; Diris y Ooghe, 2018).

Este estudio se enfoca en el mercado de créditos para educación superior financiados por el gobierno. Se estudian los determinantes del acceso a créditos educativos como método para financiar educación superior en Colombia y las asimetrías de información existentes en este mercado como el riesgo moral y la selección adversa, a través de datos nacionales de educación provistos por el Instituto Colombiano para la Evaluación de la Educación (ICFES) y el Instituto Colombiano de Crédito Educativo (ICETEX).

Desde lo conceptual, este estudio introduce un modelo de contratos con informacion asimetrcia que permite entender los resultados de este mercado. El análisis sobre el diseño de los contratos de créditos educativos ayuda en la creación de modelos de referencia como herramienta para identificar y rediseñar las funciones de beneficios y costos de ambas partes (Del Rey y Verheyden, 2009). Desde una visión de la teoría de la información se intenta construir un modelo de contrato con información asimétrica en dos fuentes: primero el riesgo moral al que se ven enfrentadas las entidades prestamistas en el momento de asignar créditos para estudiantes de lo que conocen poca información socioeconómica y financiera y ninguna sobre el esfuerzo ejercido sobre su proceso académico; y segundo la selección adversa que afrontan los estudiantes en la decisión de acceder a un crédito para financiar su educación superior o no hacerlo, dado que presentan restricciones crediticias y no logran acceder incluso al crédito (Lochner y Naranjo, 2011). El ejercicio empírico combina características de logro académico individuales con variables sociodemográficas, institucionales y financieras por medio de técnicas de análisis de *Statistical Learning* y *Machine Learning*. El objetivo es recopilar la mayor información posible sobre los individuos que solicitan créditos educativos y así poder predecir su futuro comportamiento de pago y default.

Los determinantes de incumplimiento o default de créditos educativos ha sido un tema altamente estudiado por diferentes áreas, no solamente desde una perspectiva económica (Wilms, Moore y Bolus, 1987) sino también psicológica (Flint, 1997), y se ha demostrado que el comportamiento de pago no depende únicamente de variables financieras asociadas a los ingresos de los solicitantes, sino que también las variables que miden su desempeño académico son fuertes predictores del pago del crédito (Volkwein, 1995; Flint, 1997; Monteverde, 2000; Herr y Burt, 2005). Es así como actualmente algunos de los beneficios otorgados por el gobierno para financiación de educación superior, son condicionados bajo aspectos meritorios, evidenciados a través de promedios y logros académicos para su desembolso. No obstante, no es así para las entidades privadas que ofrecen créditos educativos, ya que éstas a diferencia de las Instituciones de Educación Superior (IES), están desinformadas sobre el nivel de habilidades académicas que puedan tener los estudiantes que solicitan créditos, por lo cual el diseño de estos contratos alcanza asignaciones ineficientes en el equilibrio que afectan al estudiante o a la empresa prestamista de manera negativa.

Muchos estudios han abordado el tema de los factores que afectan el default con la ayuda de herramientas estadísticas como la econometría paramétrica, donde por lo general se han realizado estimaciones basadas en modelos logit, probit y tobit (Greene, 1989; Steiner y Teszler, 2005; Barone, 2006; Belfield, 2013) o incluso regresiones con instrumentos válidos para corregir problemas de endogeneidad con las variables referentes a las habilidades y restricciones crediticias (Carneiro y Heckman, 2002; Belfield, 2012). No obstante, pocos estudios han utilizado las técnicas de *Machine Learning*[1] para predecir los resultados de default para el sector de créditos educativos, aunque algunos estudios han analizado temas de hipotecas de vivienda, y créditos de consumo corriente (Bagherpour, 2000; Zhou y Wang, 2012). Estas técnicas requieren bases ricas en datos y covariables, por lo cual se hace también un aporte metodológico al uso de las bases del ICFES y de las técnicas de medición de default en el sector educativo.

El modelo utilizado en este caso es el algoritmo XGBoost (Chen y Guestrin, 2016), el cual a través de la calibración correcta de sus parámetros usando métodos de boostrapping y de regularización, permite el alcance de una predicción consistente evitando la sobreestimación o subestimación de modelos con alto número de observaciones y a su vez un gran número de variables independientes. En este proceso las variables con mayor relevancia para la estimación del incumplimiento del crédito a 30 días fueron las asociadas a género y educación de los padres. Sin embargo, el algoritmo permite la organización de las variables independientes por orden de importancia, es entonces que variables de contexto socioeconómico y de logro académico tomaron también parte del top 10 de las variables más importantes a la hora de predecir el default en un crédito educativo.

Los resultados del modelo empírico demuestran, al igual que Chapman (2019), que existen otras variables asociadas al incumplimiento de créditos estudiantiles que no son exactamente relacionadas con el nivel socioeconómico de los estudiantes sino también con el alcance y el desarrollo de habilidades de logros académicos. Sin embargo, haber asistido a un colegio privado y tener un auto son una proxy de nivel socioeconómico y están dentro de las 10 variables más importantes para determinar default bajo la muestra y alcance de este trabajo.

Estos resultados a su vez contribuyen a la literatura que hace hincapié en los métodos de financiación de la educación superior tanto públicos como privados, ya que conociendo cuales son estas variables claves en el riesgo de default por parte de los estudiantes que cumplen ciertas características específicas, se pueden desarrollar diseños de créditos especiales para cubrir a cada segmento de la demanda. Si bien es cierto que los créditos financiados por el gobierno como lo son los ofertados por el ICETEX en Colombia tienen un sistema de riesgo compartido entre los estudiantes y el gobierno, ya que, aunque

[1] Por ejemplo, Random Forest, Super Vectors Machine, K-Nearest Neighbours o XGBoost

existen condiciones meritorias en algunos tipos de créditos, también existen barreras crediticias como lo es la existencia de un codeudor solidario que respalde la deuda, que evitan que muchos jóvenes decidan incluso acercarse a solicitar información sobre financiación.

El resto del documento se organiza de la siguiente manera. La sección II y III se describen los datos utilizados y el contexto teórico. La sección IV presenta el modelo con información asimétrica. La sección V señala la metodología utilizada y el proceso empírico. La sección VI reporta los resultados principales y concluye con algunas sugerencias de política para el proceso de acceso a créditos para educación superior, y deja algunas ideas abiertas a discusión.

II. Contexto

A. Financiación de educación superior y acceso a créditos educativos

La demanda por educación superior se ha incrementado de manera permanente durante la última década alrededor del mundo (Ramos y Hernández, 2017), generando excesos de demanda que requieren de acciones por el lado de la oferta, tanto desde el sector público como privado. (Diris y Ooghe, 2018) usando evidencia empírica para los países miembros de la OCDE, muestran que la financiación pública a través de subsidios puede ser regresiva. Además, los retornos individuales de la educación superior ofertada por el sector privado son mayores a los del sector público. Lo anterior implica que se debe analizar de manera cuidadosa la eficiencia de la intervención de los gobiernos en el acceso a educación superior, por lo cual la financiación total por parte del gobierno es un punto que debe ser revisado minuciosamente (Salmi, 2014; Murphy, Scott y Gill, 2019).

El gasto público se asigna a ambas partes del mercado de educación superior. Desde el lado de la oferta, las IES reciben recursos con los cuales mejoran calidad en contratación de profesores, mantenimiento de la infraestructura y laboratorios etc…, por otro lado, la demanda recibe apoyo a través de becas, subsidios, beneficios de tasas de interés entre otros. Sin embargo, la demanda no alcanza a ser cubierta en su totalidad y es allí donde surgen modelos de financiación alternos como los créditos educativos, financiados con recursos privados o incluso algunos con recursos públicos, vista más como una inversión social.

En Estados Unidos los Goverment Student Loans (GSL) son créditos educativos federales, que sirven como ayuda para financiar educación superior a los jóvenes de bajos ingresos, los cuales funcionan como créditos condonables con garantía pública en caso de default (no hay riesgo para las empresas prestamistas) (Dynarski, 1991). Para Australia los Income Contigent Loans (ICL) constituyen la fuente principal de financiación de educación superior, y son un tipo de préstamo federal que a diferencia de los GLS modifican las condiciones de repago para disminuir estrés de pago y evitar default en el largo plazo (Chapman, 2006; Palacios, 2002-2011). Estos tipos de créditos son los más utilizados alrededor del mundo y aunque han tenido modificaciones de acuerdo con las dinámicas del mercado laboral y el contexto demográfico, se han mantenido hasta hoy, incluso algunos son fondeados por recursos privados de grandes empresas e instituciones privadas, que buscan inversiones a largo plazo.

Diferentes estudios y metodologías se han aplicado para la evaluación de este tipo de créditos, especialmente los financiados con recursos públicos, ya que estos son vistos como una inversión social que en el largo plazo que trae no solo rendimientos individuales sino sociales que benefician el desarrollo económico de la comunidad. A través de varios estudios para Estados Unidos, se resalta la participación de las variables de logro académico sobre el incumplimiento de pago. Entre algunos, se encuentra que el hecho de que si los prestatarios no completan la escuela secundaria implica que son más propensos a

incumplir (Dynarski, 1991; Herr y Burt, 2005), asimismo los estudiantes que se gradúan de carreras con perspectiva de salarios menores en el mercado laboral (Dynarski, 1991; Greene y Seaks, 1992). Los estudiantes con Grade Point Average (GPA) más altos o aquellos que venían con un mejor puntaje SAT (Herr y Burt, 2005) terminan siendo menos probables de incurrir en default, esto implica que existen variables académicas pre-universitarias que también influyen en el riesgo de incumplimiento (Flint, 1997; Gross et. al, 2009; Barone, 2016).

De la misma forma, las variables institucionales, familiares y demográficas complementan el análisis del riesgo de incurrir en default. Sin embargo, más allá del estado financiero del estudiante, y de las variables de ingresos individuales o familiares, su propio nivel académico y su elección del programa pueden determinar y predecir cómo será su comportamiento de pago del crédito (Palacios, 2014).

B. Crédito educativo en Colombia

El sistema de educación superior en Colombia al igual que el resto de los países de la región, ha tenido cambios significativos, no solo en la expansión de la cobertura, sino en la estructura de los programas y la creación de nuevas instituciones, las cuales se dinamizan de acuerdo con las necesidades del mercado laboral (Carranza y Ferreyra, 2019). El Gobierno Nacional ha sido un participante activo en esta dinámica a través de recursos públicos que permiten el acceso de más jóvenes de bajos ingresos al sistema de educación superior, por medio de becas de mérito y auxilios financieros, de manera que la tasa de cobertura de educación superior[2] pasó de 37,4% en 2010 a 54,3% en 2018.

Es importante resaltar que, según datos del Sistema Nacional de Información de la Educación Superior (SNIES) la tasa de cobertura de educación superior del sector privado ha pasado de 44,35% en 2009 a 50,33% en 2018 reflejando una mayor participación casi equitativa al sector público, el cual ha disminuido su participación durante los últimos 10 años. Por consiguiente, se puede pensar dos ideas, primero que los estudiantes nuevos han buscado alternativas en el sector privado cuando no logran acceder a educación superior pública por no cumplir con las condiciones socioeconómicas y/o académicas necesarias. Segundo, que los estudiantes nuevos ha se han informado un poco más sobre calidad institucional, y retornos esperados por lo cual eligen IES privadas de mejor calidad.

Una de las entidades públicas que permiten esto es el Instituto Colombiano de Crédito Educativo (ICETEX)[3], quien según el Ministerio de Educación "en el año 2002 financiaba el 9 % de los estudiantes de educación superior, hoy se financia el 19 %. Entre el 2003 y 2010 se han apoyado a 300.015 estudiantes en todas las modalidades de crédito, para lo cual se han invertido 2.6 billones de pesos" (ICETEX, 2011).

Sin embargo, es evidente que el acceso a este tipo de auxilios se ha reducido significativamente (ver figura 1), no por el hecho de que las matrículas hayan disminuido, más bien porque las maniobras públicas se limitan al presupuesto de los planes de desarrollo temporales y las medidas estructurales se ven afectadas por los constantes cambios, y para el segundo semestre del año 2019 sólo el 1,85% del total de matriculados en el sistema de educación superior fue financiado por el ICETEX (SPADIES, 2020). Esto abre paso a un tema importante en materia de acceso a educación superior que es la identificación de los estudiantes con altas habilidades que no logran beneficios públicos como becas y subsidios o acceso a este tipo de créditos respaldados por el gobierno nacional, pero que están dispuestos a financiar su educación superior a través de crédito.

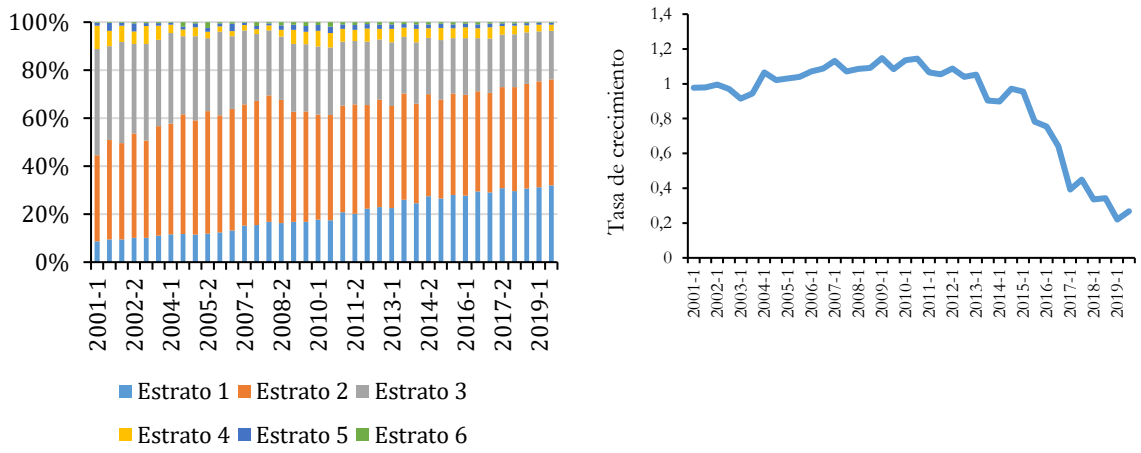
---

[2] La tasa de cobertura se toma como el cociente entre el número de individuos entre 17-21 años y el número de matriculados en pregrado de educación superior (SNIES).

[3] ICETEX fue la primera agencia colombiana en 1950 en ofertar créditos educativos en Latinoamérica. entidades financieras a la población con menores posibilidades económicas y buen desempeño académico.

Figura 1. Préstamos otorgados por ICETEX 2001-2019

a. Estrato de los estudiantes entrantes **b. Tasa de crecimiento de préstamos**



Fuente: Elaboración propia con datos de SPADIES

Los estudios realizados en materia de predicción o análisis de default en el mundo son amplios y ofrecen diferentes metodologías que fortalecen la literatura en este tópico, sin embargo, para Colombia los estudios se han concentrado en buscar opciones de créditos que puedan ser pagados de manera rápida y segura por lo cual los esfuerzos se enfocan en la variación de las tasas de interés más que analizar los componentes y un diseño apropiado para este tipo de contratos.

En Colombia, el porcentaje de estudiantes que busca financiar su educación superior a través de créditos ICETEX es bajo (ver figura 1). Sin embargo, el porcentaje que busca financiación en otras entidades financieras con fines de lucro es mayor, no obstante, no se ha tenido mucho control sobre la oferta privada de estos créditos, sino que son tratados como créditos de consumo o créditos de adquisición de activos como carros o vivienda. Es importante que se genere un tratamiento especial para este tipo de créditos, ya que involucran comportamientos de pago diferentes por parte de los estudiantes, porque al culminar sus estudios el ingreso que comenzarán a percibir por el ejercicio de su carrera en el mercado laboral, le permitirá pagar su deuda de una manera más segura. Esto es diferente a adquirir un activo como un automóvil, por ejemplo, debido a que este tipo de activos se deprecian con el tiempo, mientras que la educación eleva el valor del individuo en el mercado laboral calificado.

Es importante destacar que, aunque los créditos ofrecidos por el ICETEX tienen ventajas en cuanto a tasas de interés y flexibilizaciones en el repago para los estratos 0, 1, 2 y 3; no todos los estudiantes entrantes que vienen bajo estas condiciones socioeconómicas eligen financiar sus estudios de educación superior con esta entidad. En el panel a de la figura 1 se puede ver la distribución de los estudiantes entrantes por nivel de estrato socioeconómico en la cual se aprecia que el ingreso de estudiantes de estratos 1 y 2 viene creciendo en el tiempo mientras que los de estratos más altos no crece en la misma proporción. Esto afirma lo encontrado por (Ferreyra, Avitabile, Botero, Haimovich, y Urzúa, (2017)) quienes determinan que la demanda por educación superior, se ha transformado de manera que, en estudiantes con bajos niveles socioeconómicos y alto nivel de habilidades son quienes están abarcando la mayor parte de los cupos en instituciones de educación superior tanto públicas como privadas.

Ante esta transformación se hace necesario profundizar sobre los estudios en materia de financiación de educación superior, ya que este cambio de estructura en la demanda del mercado dio paso a otro mercado importante que es el de los préstamos educativos por parte de entidades financieras con y sin fines de

lucro y que necesita una visión panorámica de las características que pueden determinar el comportamiento de pago de un estudiante y por lo tanto, la confianza que se le podría depositar por parte de alguna entidad prestamista.

III. Datos

A. Datos y Estadísticas descriptivas

Este documento combina dos fuentes de registros administrativos del sistema de educación superior colombiano. La primera corresponde a los registros administrativos de la prueba estandarizada Saber 11 del Instituto Colombiano para la Evaluación de la Educación (ICFES), el cual recopila información a nivel individual los resultados de la evaluación de conocimientos en las áreas de lectura crítica, matemáticas, sociales y ciudadanas, ciencias naturales e inglés. Adicionalmente, contiene información socioeconómica de los estudiantes y sus hogares. Se utilizan los puntajes individuales sobre las áreas evaluadas, para aproximar el nivel de habilidades de los estudiantes de educación secundaria, además de su información socioeconómica y familiar.

La segunda fuente de información corresponde a los registros administrativos de los créditos educativos del Instituto Colombiano de Crédito Educativo (ICETEX), la cual contiene información financiera a individual sobre créditos otorgados durante los años 2015-2018 para financiar educación superior universitaria[4]. Este recurso permite un acercamiento a las características financieras de los créditos educativos, y muestra el panorama general del default de créditos públicos educativos en Colombia. También contiene información socioeconómica adicional.

Los anteriores registros administrativos se unieron usando identificadores individuales formulados por los departamentos de estadísticas del ICETEX y del ICFES quienes realizaron el pareo de los estudiantes de manera confidencial para mantener la política de tratamiento de datos y evitar la identificación personal de algún estudiante en este estudio. La base final incluye variables en cuatro categorías que afectan la probabilidad de default en los créditos educativos: individuales, familiares, institucionales y financieras. La muestra de análisis final contiene 63.753 individuos que obtuvieron al menos un crédito con el ICETEX entre 2015 y 2018. La muestra final contiene información de línea de base del formulario del ICFES y *outcomes* financieros medidos por el ICETEX. Para este análisis un individuo se encuentra en default cuando no registra pagos a la entidad prestamista luego de 30 o 60 días de la fecha de pago.

La figura 2 muestra la distribución de los créditos dentro de la muestra. El panel a presenta el tipo de crédito preferido por los solicitantes. 46,4% eligieron el crédito *Tú eliges 25%* el cual, es un crédito de largo plazo que comprende solamente estudiantes de estratos 1, 2 o 3 que cumplan con dos condiciones (i) 270 puntos en el examen nacional Saber 11 y (ii) un promedio de 3,4 en las notas de secundaria. Este crédito se caracteriza porque el estudiante solo paga el 25% del crédito durante su proceso académico y el 75% restante cuando los culmine. 12,4% eligieron el crédito *Tú eliges 30%* que es similar a *Tú eliges 25%* pero está disponible para cualquier estrato socioeconómico y el puntaje en Saber 11 requeridos de 260 puntos.

[4] ICETEX ofrece créditos para cualquier modalidad de estudios de educación superior, incluyendo pregrados, posgrados y educación técnica y tecnológica. Para fines metodológicos este trabajo solo utiliza una muestra de créditos para estudios universitarios.

Figura 2. Distribución de créditos de ICETEX

a. Tipo de crédito **b. Área del conocimiento de preferencia**

Fuente: Elaboración propia con datos de ICFES e ICETEX

Para la muestra estudiada, 23% de los estudiantes incumplieron a 30 días, 15% a 60 días, 9% a 90 días y 6% a 120 días. El ejercicio de predicción se concentra en el default a 30 y 60 días. Para ampliar el panorama general sobre los datos disponibles, se resalta que 58% de las observaciones corresponden a mujeres y 42% a hombres. Este dato muestra que la incursión de la mujer en la educación superior ha tenido un crecimiento importante que deja sobre la mesa de debate la discusión sobre las opciones de permanencia de las mujeres en el sistema educativo quienes son las que mayormente desertan según el Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES) y las que mayormente incumplen (57% de las mujeres incurrieron en default a 30 días) pero al parecer son quienes más buscan opciones de financiación.

IV. MODELO CON INFORMACIÓN ASIMÉTRICA

Para identificar fallos o asimetrías de información dentro de cualquier mercado de manera eficiente, es necesario conocer cómo deberían funcionar estos mercados de manera exitosa. Por lo tanto, se propone un modelo de referencia con información perfecta y completa y luego el modelo en equilibrio competitivo, pero con asimetrías de información. (Bardhan y Udry, 1999) Desarrollan cuatro modelos de referencia para créditos agrícolas, basados en un enfoque de juegos bayesianos con información asimétrica. Sin embargo, dejan abierta a discusión la idea de adaptarlos para créditos educativos mas no la desarrollan, y ese es el propósito de esta sección.

Modelo de referencia con información perfecta y completa

Supuestos:

- Existe una empresa prestamista y un estudiante, ambos agentes representativos.
- Todos los agentes de este mercado son neutrales al riesgo (bajo información perfecta y completa).
- Los estudiantes tienen acceso a la misma oferta institucional de programas universitarios.
- El conjunto de estrategias de las empresas se basa en la oferta de un número finito de préstamos.
- El conjunto de estrategias de los estudiantes es elegir un préstamo del menú finito que ofertan las empresas $[0, \infty]$ o pueden elegir ninguno.
- Para este caso donde todos conocen la información disponible y hacen uso eficiente de ella, la elección es simultánea.
- Ambos tienen costos fijos iguales a 1 ($C = 1$).

- Los estudiantes obtienen un salario básico (W_b) si no pueden acceder a educación superior, de lo contrario obtienen un salario forma (W_{es}), donde: $W_{es} > W_b \geq 0$
- Sea $\pi(e)$ la probabilidad de que el estudiante salga exitoso del programa universitario que eligió. Este es un índice que mide el esfuerzo que el estudiante decide poner sobre su proceso educativo.
- Sea $D(e)$ el costo de ingresar a educación superior (costos de matrícula, transporte y de bolsillo)
- Las empresas prestamistas imponen un factor de interés i tal que: $i \leq W_{es}$
- Las empresas prestamistas tienen acceso a un mercado de capitales alternativo al de créditos educativos, donde obtienen un retorno de ρ $(W_{es} > \rho \geq 1)$.
- Si el estudiante es exitoso, no puede renunciar a su promesa de pagar el crédito a la empresa prestamista.
- Los retornos de la empresa y de los estudiantes se pueden estructurar de la siguiente manera:

Tabla 1. Pagos de los agentes

Logro Académico (Educación superior)	*Estudiante*	*Empresa Prestamista*
Exitoso	$W_{es} - i - D(e)$	i
No exitoso	$D(e)$	*0*

Elaboración propia, basada en (Bardhan y Udry, 1999)

De esta manera, cuando el estudiante logra terminar de manera exitosa su educación superior obtiene un salario mayor al que obtendría sin ella, además paga a la empresa los intereses generados y recibe un beneficio neto luego de los costos incurridos. En caso de no lograr exitosamente la culminación del proceso educativo el estudiante obtiene una desutilidad expresada a través de los costos y la empresa prestamista obtiene cero, en lugar de pérdidas, esto dado que ambos tienen los mismos costos fijos, y la empresa no presenta en este caso costos adicionales como estudios de créditos entre otros.

- Se puede entonces determinar que la utilidad esperada del estudiante y los beneficios esperados de la empresa son de la siguiente manera:

$$UE(i,e) = \pi(e)(W_{es} - i) - D(e)$$
$$\pi(i,e) = \pi(e) \cdot i$$

Equilibrio competitivo con información completa:

El desarrollo de la interacción hacia el equilibrio es el siguiente:

- Las empresas prestamistas ofrecen un menú de créditos educativos con un rango de cobertura monetaria fija.
- Dado que hay información perfecta y completa, las empresas prestamistas observan (e).
- Los estudiantes eligen un crédito de la lista si y solo si este les genera un retorno mayor o igual a su utilidad esperada.
- Existe entonces un par de (i_1, e_1)que solucionan las siguientes restricciones:
  a) $UE(i_1, e_1) \geq W_b$
  b) $\pi(i_1, e_1) \geq \rho$
  c) No existe otro par de (i, e) que ofrezca un retorno mayor o igual a ρ
- Dadas las condiciones anteriores, el problema del estudiante representativo es:

$$Max \qquad \pi(e)(W_{es} - i) - D(e)$$

$$S.a \qquad \pi(e) \cdot i \geq \rho$$

$$\pi(e)(W_{es} - i) - D(e) \geq W_b$$

Así, para cualquier valor de ρ existe un W_b suficientemente bajo para que ambas restricciones se cumplan. De esta manera en el equilibrio se cumple que:

1. $Cmg = Img$
2. $\pi(e_1) \cdot i_1 = \rho$
3. $U(i_1, e_1) = \pi(e_1) \cdot W_{es} - D(e_1) - \rho > W_b$

En el equilibrio competitivo con información perfecta y completa, las condiciones de optimalidad se cumplen para ambos agentes, de manera que aun estos maximizando su utilidad y su beneficio respectivamente de manera independiente, alcanzan un equilibrio en el cual la empresa prestamista obtiene un valor igual al que obtendría en otra inversión alternativa, y el estudiante por su parte, obtiene un beneficio mayor que el salario básico que obtendría si no accede a educación superior, aun descontando costos e intereses pagados a la empresa.

A. Equilibrio competitivo con riesgo moral y selección adversa

El equilibrio competitivo con información perfecta y completa parece comportarse bien de acuerdo las intenciones de cada agente, sin embargo, en la realidad, los préstamos en general contienen factores de riesgos como el no pago por parte de los estudiantes a las empresas (Chapman, 2016), o que el estudiante elija un programa profesional con información incompleta sobre sus retornos (Bernal, 2019) y no logre éxito académico al final del periodo.

Por lo anterior, se presenta ahora un modelo con información asimétrica que hace hincapié en el riesgo moral y la selección adversa, esto para entender un panorama más real sobre los riesgos que afrontan los contratos de créditos educativos actualmente.

Los estudiantes por su parte, al no informarse completamente sobre retornos, costos de matrícula, condiciones de persistencia entre otros factores, representan un riesgo para la empresa prestamista, por lo cual, éstas no tienen la capacidad de diferenciar si el estudiante es riesgoso o no lo es dado su nivel de habilidades académicas, por lo cual, incluyen dentro de su sistema de creencias, que existen dos tipos de estudiantes: los riesgosos y los no riesgosos.

Las empresas se afrontan a un problema de riesgo moral implícito en el hecho de que realmente, no conocen el nivel de esfuerzo ejercido por los estudiantes sobre su proceso educativo, y por lo tanto se deben crear conjeturas o creencias que le informen un poco más acerca de a qué tipo de estudiante se enfrentarán. Estas creencias conllevan a una nueva restricción sobre el problema inicial, ya que las empresas deben ofertar un menú de créditos educativos, tal que la elección del estudiante sobre el crédito induzca a la empresa sobre el tipo del estudiante. Sin embargo, las empresas prestamistas no se arriesgarán a perder su inversión a cero costos, como en el modelo anterior, por el contrario, utilizan una herramienta financiera para aliviar estos riesgos, generalmente es el uso de un colateral (C).

Las empresas, solicitan que se cuente con una garantía adicional, puede ser una persona natural que demuestre ingresos suficientes para respaldar la deuda en caso de no pago por parte del estudiante, incluso en ocasiones se transan los derechos de propiedades privadas como parte del contrato. Estos colaterales, constituyen uno de los principales problemas de acceso a créditos educativos (Lochner y Naranjo, 2011; Chapman, 2009; Solís, 2002). En este caso el colateral está en función del esfuerzo del estudiante, ya que no solo depende de una cantidad monetaria sino del logro académico de los estudiantes en el desarrollo

de su proceso educativo, que se puede extraer de las variables académicas extraídas del examen SABER 11 del ICFES. El modelo se desarrolla de la siguiente manera:

Modelo:

- Ambos agentes se vuelven aversos al riesgo.
- Las empresas prestamistas no observan el nivel de esfuerzo, pero pueden informarse sobre el tipo de estudiante ofertando créditos diferentes para cada tipo.
- En este caso $t = \{1,2\}$ determinará el tipo de estudiante tal que $t = 1$ determina si el estudiante se considera "un deudor cumplido" y $t = 2$ si el estudiante es considerado "un deudor incumplido".
- El conjunto de estrategias de cada jugador es el siguiente: los estudiantes eligen un préstamo de un menú finito que oferta la empresa $[0, \infty]$ o pueden elegir ninguno. Para las empresas prestamistas el conjunto de estrategias sigue siendo ofertar un menú de préstamos segmentados a los estudiantes de acuerdo con las características observables de los mismos.
- Por lo anterior, t puede inducir el conocimiento de la empresa sobre el nivel de esfuerzo e, por lo tanto $t = 1$ lleva a la empresa a creer que el estudiante es de altas habilidades académicas.
- Dado que un tipo es más riesgoso que otro, los retornos para ambos también son diferentes, de modo que: $\pi(1) > \pi(2)$, entonces $\pi(t)W_{es}(t) = W_{es}\ \forall\ t$
- Existe una nueva restricción al problema anterior, tal que, el estudiante elija cierto crédito asociado a su nivel de habilidades, de modo que se cumpla que esa elección le generará una utilidad esperada mayor que cualquier otro crédito, dadas sus características individuales.
- La utilidad esperada del estudiante es $u(i,t) = \pi(t)[W_{es}(t) - i(t)]$
- El beneficio esperado de la empresa depende de la tasa de interés y del tipo de estudiante: $i_1 < i < i_2 \quad y \quad \pi(i,t) = \pi(t) \cdot i$
- Existe una función $C(t)$ que induce al estudiante a poner todo su esfuerzo al menos uno muy alto en su proceso académico.
- Las decisiones se toman de manera secuencial.
- El problema para cada tipo de estudiante se convierte entonces en un problema de dos agentes con incertidumbre asociada a sus pagos. Los pagos, de cada agente se maximizan de acuerdo con sus funciones de distribución de probabilidad que representan aversión al riesgo:

$$\begin{aligned} Max_{\{e,i,c\}} \quad & \pi(t)[W_{es}(t) - i(t)] - [1 - \pi(t)] \cdot C(t) - D(t) \\ S.a \quad & \pi(t) \cdot i(t) + [1 - \pi(t)] \cdot C(t) \geq \rho \\ & \pi(t)[W_{es}(t) - i(t)] - [1 - \pi(t)] \cdot C(t) - D(t) \cdot C(t) - D(t) \geq W_b \end{aligned}$$

$$\pi(e)(W_{es} - i) - [1 - \pi(e)] \cdot C(e) - D(e) \geq \pi(e')(W_{es}(e') - i) - [1 - \pi(e')] \cdot C - D(e')$$

De esta manera en el equilibrio se debe cumplir que:

1. $C = i = \rho$ Esto indica que el riesgo es compartido entre ambos agentes. Por lo general el riesgo es transferido completamente al prestamista (estudiante), dado que el colateral es expresado como una porción del salario de otra persona o de un activo. Sin embargo, en este caso el colateral depende también del tipo del estudiante, tal que, si éste es exitoso académicamente representará un riesgo menor para la empresa ya que al entrar al mercado laboral con un salario W_{es} mayor puede realizar el pago oportuno de su deuda.

2. $e^* = e_t$ El índice de esfuerzo es independiente de cada tipo en el equilibrio y estará determinado por el puntaje de logro académico obtenido en este trabajo en la siguiente sección con herramientas econométricas.
3. $u(i_t(t), t) \geq W_b$ La utilidad esperada del estudiante en función de su tipo y de la tasa de interés impuesta por la empresa debe ser al menos igual al salario básico que obtendría en otro empleo sin educación superior.
4. $\pi(i_t(t), t) \geq \rho$
5. $u(i_1, t) \geq u(i_2, t)$ esto indica que la utilidad esperada del estudiante riesgoso es mayor que la del no riesgoso, dadas las condiciones diferenciadas de tasas de interés y esfuerzo académico.
6. $W_{es} - \rho > W_b$, de otra manera, ningún estudiante recibe créditos.
7. No existe una tasa de interés $i(t)$ que genere retornos mayores o iguales a ρ
8. El comportamiento de la tasa de interés es primordial, porque a medida que aumente i^* los estudiantes de tipo no riesgoso van a abandonar el mercado y por lo tanto el beneficio del prestamista empezará a caer.

Sin embargo, el equilibrio competitivo sin información asimétrica genera mayor nivel de beneficio y de utilidad a la empresa y al estudiante respectivamente, por lo tanto, uno de los objetivos clave en el diseño de un contrato de crédito educativo es reducir a través del uso de la información disponible las asimetrías de información existentes, de modo que se habiliten nuevos conjuntos de información que permitan la inclusión de nuevas estrategias y acciones por parte de los agentes, que sean alcanzadas en el equilibrio.

Es importante resaltar que, aunque se minimizan las asimetrías de información, aún existen mejoras que pueden hacer que alguno de los individuos se desvíe del equilibrio competitivo, ya que, en la realidad, no existe un menú pequeño de créditos, sino una amplia gama de ellos, con diferentes beneficios y costos.

V. METODOLOGÍA (ESTRATEGIA EMPÍRICA)

A. Un primer modelo

El objetivo es construir un modelo que permita estimar la probabilidad de incurrir en default de un estudiante, condicionado a sus características académicas, económicas y financieras. De manera que, más adelante este modelo permita la flexibilización de las medidas de acceso a créditos de este tipo, y así a la ampliación de la cobertura de educación superior (Dynarsky, 1991,1994; Greene y Seaks 1991).

Habitualmente, se pensaría que un modelo logit multivariable (Carneiro y Heckman, 2002; Steiner y Teszler, 2005) o incluso un modelo tobit (Greene, 1989) bastaría para estimar la variable dicotómica de default y que utilizando variables instrumentales de logro académico se lograría llegar a un coeficiente consistente y eficiente. Sin embargo, los supuestos de normalidad y teoría asintótica que se requieren aplicar para esto pueden dilatar la forma real de la función de distribución de probabilidad de los datos, y puede generar estimaciones sesgadas.

Para este caso se plantea inicialmente un modelo común para este tipo de análisis que es la regresión logística con varias variables. Este modelo se estima por máxima verosimilitud suponiendo que la función de distribución de probabilidad distribuye logística y cumple con los supuestos clásicos del modelo de probabilidad. Se plantea al siguiente modelo:

$$\Pr(d = 1|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{F}) = \varphi(\beta_0 + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{Y} + \beta_3 \boldsymbol{F})$$

Donde $d = 1$ señala que el estudiante se encuentra en default de X días (ej. X=30), y la estimación por Máxima verosimilitud (MVL) genera los efectos parciales de cada variable en cada vector y así identificar

qué variables inciden de una manera significativa sobre el estado de default del estudiante. Una vez estimados los parámetros asociados a este modelo, se estima la predicción de default.

No obstante, los modelos logit tienen ciertas limitaciones el proceso de predicción que han sido documentadas en diferentes ocasiones aproximando su resultado a uno menos eficiente y predatorio que los algoritmos de predicción de *Machine Learning* como *Random Forest* (Zhao, Yan, Yu, & Hentenryck, 2020)[5]. Los autores demuestran que los algoritmos basados en gradientes descendientes como lo son los árboles de decisión pueden alcanzar a capturar relaciones no lineales existentes entre las variables independientes que en un modelo logit pueden causar violación a los supuestos sobre el comportamiento de la función de probabilidad conjunta y generar estimaciones menos confiables. Una vez estimado este modelo, se compara su alcance con las estimaciones del modelo no paramétrico y así evaluar la diferencia en los errores de clasificación de ambos métodos.

A. Algoritmo XGBoost

Desde el enfoque de aprendizaje supervisado del *Machine Learning* se han desarrollado herramientas innovadoras ajustadas al entorno computacional para resolver problemas de clasificación y pronóstico en grandes muestras de forma más eficiente. Una de estas herramientas es el algoritmo *Gradient Boosting Decision Tree* (*GBDT*), el cual combina dos componentes que fortalecen las estimaciones débiles iniciales. El *GBDT* estima inicialmente árboles de decisión con poca significancia llamados clasificadores débiles o *stumps*, pero a través de métodos de *ensembling*, como el *Boosting*, los árboles inicialmente estimados son ajustados minimizando errores en cada iteración que los convierten en fuertes clasificadores con mejor información.

Tabla 2. Definición y explicación de variables

Símbolo	*Significado*
X	Covariables, todas las variables explicativas
Y	Variable dependiente: default (si=1, no=0)
y_i	Valor observado de default para el individuo i
$\widehat{y_i}^t$	Predicción de default para el individuo i
$f_m(x)$	Output value
$F_m(X)$	m clasificadores débiles que fueron añadidos al modelo final predecible con un cierto peso después de m veces iteración
$L(y_i, \widehat{y_i})$	Función de pérdida individual
$L(F_m(X), Y)$	Función de pérdida total
γ	Parámetro de penalización
T	Numero de nodos terminales
λ	Parámetro de regularización
η	Tasa de aprendizaje (Learning rate)
$L_m(f)$	Función objetivo
$P(f_m)$	Regularización

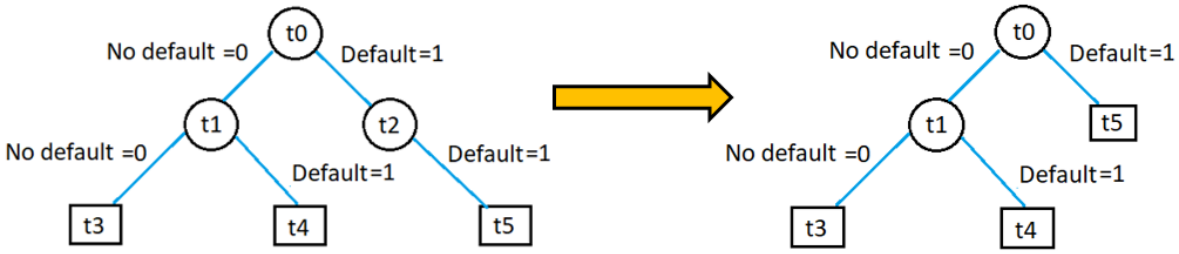
Fuente: Elaboración propia basada en (Ma., Sha, Wang, Yu, Yang & Niu, 2018)

Los árboles de decisión son creados a partir de la data original y no requieren supuestos funcionales sobre la misma, por esto se dice que es un enfoque no paramétrico de estimación. Estos árboles particionan recursivamente los datos en diferentes regiones asociando las observaciones a través de la evaluación de

[5] Los autores también generan resultados para comparaciones entre XGBoost y LightBMM

los residuales al cuadrado (SCE) que arroja la estimación de cada árbol y así va seleccionando los nodos siguientes. Los árboles se van "podando" de modo que a medida que se ajusta para la siguiente estimación, como se observa en la figura3, de manera que, elimina las ramas o aquellos cortes que tienen mayor error de clasificación y así determina una extensión adecuada del árbol. La elección de cuántos nodos terminales T pueda tener el árbol, se puede determinar mediante *Cross-Validation (CV)*, para evitar que árboles muy grandes sobreestimen el modelo o algunos muy pequeños no alcancen a capturar la estructura real de los datos.

Figura 3. Creación y poda de árboles de decisión.

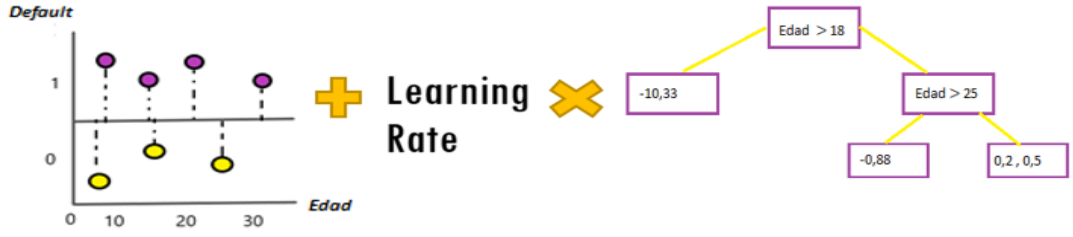


Fuente: Elaboración propia basada en Louppe(2014)

El algoritmo *eXtreme Gradient Boosting* (XGBoost), que está basado en la estructura de GBDT (Chen y Guestrin, 2016), pero que a diferencia de los algoritmos existentes como Pgbrt, Sklearn y R.GBM, está diseñado para alcanzar una estimación del modelo más veloz y eficiente con relación a los errores de clasificación, de manera que reduce la complejidad computacional en grandes muestras, especialmente para pronósticos y estimaciones de variables dicótomas.

Específicamente XGBoost minimiza la función de pérdida de los árboles de decisión completos, a través de la calibración de los parámetros presentados en la tabla 2, a diferencia de las *Random Forest*, las cuales ajustan iterativamente *stumps* o árboles relativamente cortos (que tienen solo dos hojas o nodos terminales), aunque también los ajusta iterativamente, XBoost alcanza de manera más rápida la predicción más cercana al objetivo, gracias al ajuste óptimo de sus parámetros de regularización (*regularization parameters*) y tasas de aprendizaje (*Learning rates*), que permite escalar los árboles a través del *Boosting* y crear nuevos minimizando la función de pérdida del árbol anterior, de modo que corrige en cada iteración el error de clasificación, concluyendo en un resultado consistente y robusto a datos heterogéneos.

Figura 4. Proceso de optimización del algoritmo *XGBoost.*



Fuente: *Elaboración propia.*

XGboost busca minimizar la función de perdida de cada observación para crear árboles de decisión que resuman la estructura de los datos en regiones, de manera que crea clasificadores débiles y a través de

Boostraping, logra que los árboles más nuevos tengan menores errores de predicción que los anteriores, entonces los últimos árboles estimados se vuelven en clasificadores fuertes. En este caso la función de perdida es el negativo de la función de log-verosimilitud para especificar que se trata de un problema de clasificación:

$$L(y_i, \hat{y}_i) = -[y_i Log(\hat{y}_i) + (1 - y_i) Log(1 - \hat{y}_i)]$$

Por lo tanto, el algoritmo utiliza las funciones de pérdida individual para minimizar la siguiente función:

$$L_m(f) = \Sigma_{i=1}^{n} L(y_i, \hat{y}_i) + \Sigma_{t=1}^{T} P(f_m)$$

$$L_m(f) = \Sigma_{i=1}^{n} L(y_i, \hat{y}_i) + \gamma T + \frac{1}{2}\lambda f_{m_i}$$

Donde $P(f_m) = \gamma T + \frac{1}{2}\,\lambda\,\Sigma_{t=1}^{T}\, f_m$ es el término de regularización del modelo, el cual permite que el modelo no se afecte significativamente por observaciones atípicas, y así evitar el fenómeno común de sobreestimación en muestras grandes. La meta es encontrar f_m de cada hoja del árbol que minimice toda la ecuación de pérdida total, este valor es el cociente de la suma de los residuales al cuadrado (SCE) y la suma del número de residuales más lambda:

$$f_m = \frac{\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n} [\hat{y}_i (1 - \hat{y}_i) + \lambda]}$$

Específicamente el algoritmo utiliza una aproximación de Taylor polinomial de segundo orden para estimar este valor y hacer crecer el árbol a través del cálculo del puntaje de similitud (SS) de cada hoja. Por último, elimina aquellos cortes que van incrementando su error de clasificación[6] y los reajusta a las hojas ya existentes. De este modo, las siguientes iteraciones serán menos sesgadas.

$$SS = \frac{\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n} [Probabilidad\ anterior_i \cdot (1 - Probabilidad\ anterior_i)] + \lambda}$$

Para elaborar el modelo no paramétrico, se toma aleatoriamente una porción de la data original y será la base de datos de entrenamiento y el resto de los datos serán parte de la base de prueba. Sobre la data de entrenamiento se realizará todo el proceso de construcción, modificación y elección de variables y parámetros más relevantes y sobre la data de prueba evaluaremos el modelo resultante.

En la Tabla 3 se especifican los parámetros que serán utilizados en el proceso de construcción, estimación y evaluación del modelo y su significado:

Tabla 3. Parámetros del algoritmo XGBoost

Parámetro	*Significado*
Objetivo	Función objetivo a utilizar.
Rmse	Error cuadrático medio
Logloss	Función del negativo de log-verosimilitud
error:	tasa de error de clasificación: para este caso cada predicción mayor a 0.5 será $y = 1$ de otra forma será $y = 0$
AUC	Área bajo la curva de la evaluación de la clasificación

[6] La ganancia o *gain* de cada árbol se calcula teniendo en cuenta el puntaje de similitud o *similarity score.*

Eta	Parámetro de regularización que escala la nueva predicción, de manera que evita el sobreajuste y es un valor $[0,1]$. En este caso $eta = 0.3$
max_depth	Máxima profundidad del árbol
min_child_weight	Hace referencia al número mínimo de muestras necesarias para establecer cada modelo. El rango de valores es [0, ∞].
Nrounds	Número de iteraciones, en este caso 1000

Fuente: Elaboración propia.

Finalmente, la ventaja del algoritmo es que reduce el tiempo y la velocidad de estimación computacional de manera que, utiliza paralelización en grandes muestras sin necesidad de aplicar complejos conceptos teóricos para simplificar la medición, además reduce la complejidad de los árboles de decisión, por lo cual, el proceso de podado es más eficiente. Por último, XGBoost utiliza los modelos ya estimados para hacer nuevas estimaciones, de esta manera evita la pérdida de información sobre los errores ya estimados.

La calibración correcta de los parámetros de regularización y de ajuste iterativo de los árboles de decisión, se realiza a través de *Cross – validation*. Esta técnica permite particionar la data en 10 o 5 grandes bloques de información y a través de boostrapping encuentra el conjunto de parámetros que genera la estimación con el menor error de entrenamiento posible. La evaluación del modelo estimado se da a partir del puntaje de precisión (*Accuracy*) y del alcance en el área bajo la curva, en este caso en cuanto la predicción se acerque a 1 su alcance es mejor.

Algo sobresaliente de los algoritmos no paramétricos de predicción es que permiten la organización de las variables en orden de importancia. Esto a su vez permite identificar cuál es la variable que más participación tiene sobre la predicción y cuál es la que menor peso tiene.

VI. RESULTADOS

En esta sección se presentan los principales resultados de este documento. El método de estimación principal es el del aprendizaje automático o *Machine Learning* para problemas de clasificación, en particular el método XGBoost. Los modelos paramétricos normalmente no capturan la compleja estructura de los datos, los cuales cuentan con relaciones no lineales entre las variables que pueden afectar la estimación por máxima verosimilitud que utilizan, por ejemplo, los modelos logit. No obstante, en el anexo A.1 se presentan y discuten los resultados generados por el modelo paramétrico que asume distribución logística.

A diferencia de los modelos paramétricos, los modelos XGBoost son completamente flexibles y son compatibles con cualquier tipo de no linealidades en los datos. Para el ejercicio empírico se aplicaron los valores para los parámetros del algoritmo XGBoost señalados en la tabla 3. Estos parámetros fueron seleccionados mediante Cross – validation con 1000 boostraps. Adicionalmente, y dado que el algoritmo XGBoost en más eficiente computacionalmente bajo variables categóricas, en este ejercicio se convirtieron todas las varaibles categóricas como nivel de estrato, educacion de los padres, tipo de colegio, área de conocimiento etc, en variables binarias de tal forma que el algortimo pueda paralelizar las matriz de variables independientes con la matriz del conjunto de parámetros de manera más rápida.
A diferencia de los modelos paramétricos, los modelos XGBoost son completamente flexibles y son compatibles con cualquier tipo de no linealidades en los datos. Para el ejercicio empírico estas características son importantes, ya que las limitaciones sobre el conocimiento profundo de los datos disponibles y los métodos frecuentistas pueden restringir el análisis de relaciones entre los factores que

podrían explicar de manera más precisa el problema central y que son omitidos por la no linealidad del mismo conjunto de datos.

A pesar de ser un modelo altamente demandado en el análisis de big data, el modelo XGBoost presenta limitaciones ante la existencia de muchos valores faltantes y debe ser calibrado de acuerdo a las características del conjunto de datos. Lo anterior causa deficiencias en la estimacion, ya que los árboles de decisión se construyen a partir de la informacion disponible para cada individuo de manera que divide por orden de relevancia el conjunto de características disponibles, que pueden afectar el indice de similitud y de sensibilidad del modelo. Una ventaja es su eficiencia computacial que si bien es rápida y precisa, puede mejorar su prediccion con la estimacion de los modelos en computadoras con alto grado de procesamiento de paralelizacion de datos.

Para la estimacion correcta del modelo, XGBoost mejora su prediccion cuando todas las variables independientes refieren categorías, por lo tanto, se procedó a convertir cada categoría de los factores disponibles de las bases de ICFES e ICETEX en variables binarias. La lista de variables utilizadas se puede ver en la tabla A.2 de los anexos y el valor de los parámetros se encuentra en la tabla 4.

Tabla 4. Parámetros del algoritmo XGBoost utilizados

Parámetro	*Valor*
Función objetivo	binaria logística
Error_metrics:	error
Eta	1
max_depth	10
min_child_weight	10
Nrounds	1000
gamma	0.5
Alpha	0.1
Nfold	10

Elaboración propia

Para el alcance de este trabajo se convitieron en variables binarias todos los factores, pero para los asociados al logro académico se tomaron en cuenta escalas de puntaje, de manera que el puntaje de las pruebas individuales de matemáticas, lectura crítica e inglés se dividieron en bloques de puntajes, el primer bloque corresponde a la obtención de 0-25 puntos, el segundo bloque 25.1-50 puntos y así sucesivamente hasta completar 4 bloques o intervalos de puntos la variable binaria está asociada a qué intervalo de puntaje obtuvo el estudiante en el examen Saber 11.

Tabla 5. Evaluación del modelo

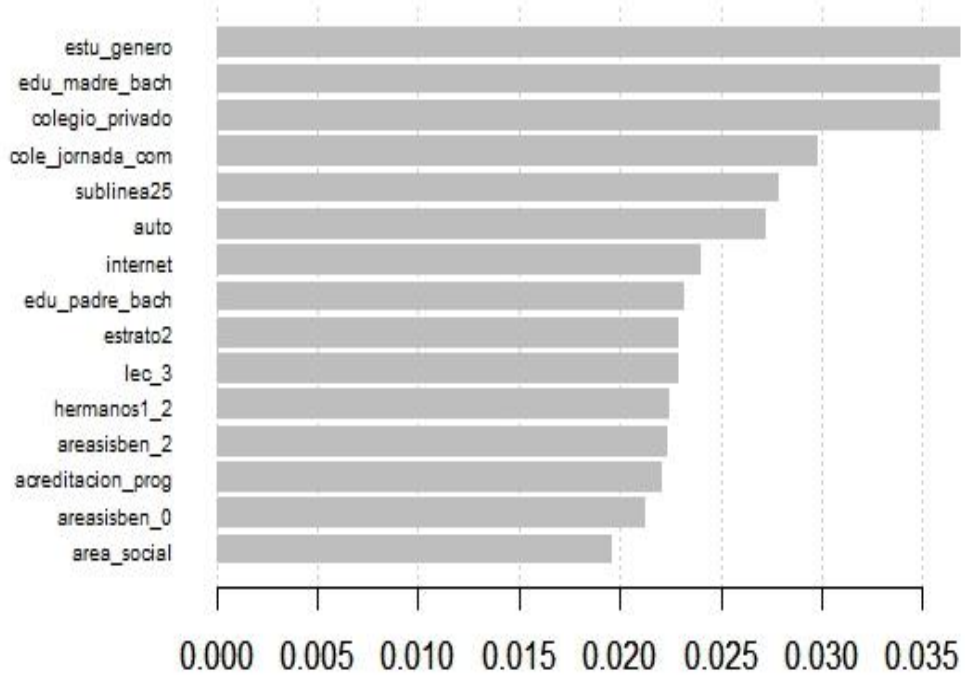
Variable	*Resultado*
Presición (Accuracy)	79%
Área bajo la curva	51%
Error de entrenamiento	0.002420
Log-loss de entrnamiento	0.041336
Error de prueba	0.02075
Especificidad (Specificity)	91%

Elaboración propia

La evaluacion del modelo XGBoost se mide de acuerdo al alcance de su precisión, tanto para los casos positivos como para los negativos. La tabla 5 muestra que el modelo pudo predecir el 79% de los casos positivos de default con un área bajo la curva de 51% y un error de entrenamiento menor que de el prueba. Esto último se debe a que se tomó 70% de la data para entrenamiento y el 30% restante para pruebas de ajuste del modelo.

La figura 3 muestra el top 15 de las variables que más contribuyen al modelo, del total de variables disponibles mostradas en la tabla A.1. Los factores que más contribuyen (comenzando por el más importante) son el género del estudiante, la educación de la madre (si tiene al menos bachillerato), si asistió a un colegio privado, si el colegio era de jornada completa, el tipo de crédito en este caso *Tú eliges 25%,* tener automóvil, tener acceso a internet, la educacion del padre (si tiene al menos el bachillerato), nivel estrato 2, puntaje de lectura entre 51.1-75, el número de hermanos entre 1 y 2, si pertenece al sibén de área 2, si el programa de educacion superior que desarrolla está acreditado en alta calidad y si su área de conocimiento es de ciencias sociales y humanas.

Figura 3. Top 15 de las variables más importantes



Elaboración propia a partir de resultados

Aunque las variables sociodemográficas son los principales predictores en el momento de evaluar el default, las habilidades académicas, aproximadas por el puntaje en las pruebas individuales evaluadas en el examen Saber 11 y las variables del colegio, también toman importancia en explicar el comportamiento de pago de los estudiantes. Cabe resaltar que el tema de género sigue siendo un campo importante de investigación, ya que a pesar de los avances en materia de equidad han sido sobresalientes, los efectos de ciertas políticas o cambios aplicados a la sociedad aún siguen afectando de manera destina a hombres y mujeres, en este caso ser mujer es un fuerte predictor del incumplimiento de los préstamos estudiantiles,

toda vez que en trabajos como (Miller, 2017) se exponen las dificultades que enfrentan las mujeres en torno a los préstamos estudiantiles, en cuanto a las razones de deserción y el entorno laboral poco flexible al que se enfrentan al culminar sus estudios.

El capital humano del hogar juega un rol fundamental en el default. La educación de la madre y el padre sigue siendo un tema relevante a la hora de analizar tópicos financieros alrededor de los hijos. El género y la educación de los padres explican en conjunto alrededor del 9% del modelo. Especialmente en este caso bajo los requerimientos para solicitar y obtener créditos del ICETEX se debe presentar una persona responsable en caso de no pago por parte del estudiante. 40% de los estudiantes presentaron un deudor solidario, 13% presentó su madre como deudor solidario, 13% a alguno de sus padres y el 34% restante presentó otro deudor como abuelos, abuelas, tíos, tías entre otros.

Adicionalmente, tener un automóvil y acceso a internet son fuertes predictores y que además está relacionados con el nivel económico de los estudiantes. Esto fortalece el hecho de que el comportamiento de pago sigue estando estrechamente relacionado con el contexto socioeconómico del estudiante. En Colombia, tener un auto en el hogar representa en parte la capacidad de endeudamiento de las familias, adicionalmente el acceso a internet también constituye un factor esencial para la estimación de diferentes análisis. (Osunade, Ojo & Ahisu, 2009) encontraron que estudiantes con acceso a internet en Nigeria, tenían mejores resultados en el logro académico, que aquellos que no tenían acceso a internet. Es importante analizar estos resultados para generar argumentos válidos de calidad para el diseño e implementación de políticas sociales que permitan que el proceso de tejido social se solidifique a través del alcance de las tecnologías para todos los estudiantes.

Así mismo, el asistir a un colegio privado constituye un fuerte predictor del comportamiento de pago de los estudiantes, toda vez que el hecho de pagar cierto monto de pensión para el colegio, puede decir que existen esfuerzos no solamente individuales sino familiares hacia la inversión en educación y que fortalece la idea de que la familia puede garantizar el pago del crédito en caso de no pago por parte del estudiante.

Para este caso finalmente el alcance del logro educativo y habilidades se ven reflejadas en la variable dentro del Top que fue el puntaje de lectura crítica (2%). Este resultado respalda los encontrados por (Chapman, 2014), y (Herr y Bur, 2005) en sus respectivos trabajos, en los cuales variables de logro académico como puntajes en exámenes nacionales son fuertes predictores del default en los préstamos.

Se resalta entonces, la necesidad de la evolucionar el diseño y evaluación de los préstamos estudiantiles como financiamiento de educación superior, de manera que se amplíe la visión panorámica acerca de los incentivos y costos para cada parte del contrato y así generar mejores resultados en el equilibrio.

CONCLUSIÓN

A lo largo de este trabajo, se ha hecho hincapié sobre la necesidad de ampliar la cobertura de educación superior a través de créditos educativos, para solventar a la creciente demanda (Ferreyra y Carranza, 2017) y sobre la importancia de diseñar de manera eficiente los contratos asociados a este tipo de préstamos.

Como se mencionó en el marco referencial, es de saber común que las características socioeconómicas y demográficas de los estudiantes determinan el comportamiento de pago bajo estándares netamente financieros los cuales muchos potenciales estudiantes no alcanzan a cumplir, limitando así su acceso al sistema de educación superior. Sin embargo, también se refirieron aportes investigativos de países como Estados Unidos, China y Australia sobre el

hecho de que las habilidades de logro académico aportan al perfilamiento de un estudiante ante una agencia prestamista, de manera que le brinda más información que internalizan en los contratos de préstamos educativos y así comparten con mayor confianza los riesgos asociados al incumplimiento de pago.

(Chapman y (2014), encuentra que la probabilidad de acceder al mercado laboral incrementa al tener éxito académico medido por factores como; graduarse dentro del tiempo esperado, no incumplir los créditos dispuestos en cada período académico, entre otros. Debido al éxito de estos contratos, los modelos de predicción de default han evolucionado con el tiempo y con el alcance de grandes cantidades de información (*Big data*) se han podido refinar algoritmos eficientes que se utilizan de manera común en el campo financiero.

En Colombia el ICETEX brinda créditos educativos alternos a los ofertados por otras entidades financieras, toda vez que enfoca sus recursos en la población de menores ingresos. No obstante, muchos estudiantes no solicitan préstamos bajo estas modalidades por encontrar barreras crediticias asociadas a su contexto socioeconómico.

Como resultado de la adaptación de un algoritmo de *Machine Learning* bajo un enfoque supervisado para problemas de predicción de default en créditos educativos, se tienen resultados que alimentan la literatura sobre los factores que determinan el comportamiento de pago de un estudiante que cumple con ciertas características. Gracias a los datos provistos por el ICETEX y el ICFES, se pudo estimar que dentro de los 15 principales factores existen variables asociadas al entorno familiar, al contexto socioeconómico, financiero y de logro académico.

Actualmente el acceso a la información de grandes muestras ha permitido el desarrollo de herramientas que contribuyen a la construcción de modelos económicos e incluso de carácter social que permiten un mejor alcance en términos de consistencia y precisión de las políticas económicas en diferentes sectores. Para este caso, la educación como clave del desarrollo económico y social debe permanecer en la agenda política de la administración pública, de manera que la gestión de los recursos destinados al crédito educativo en Colombia, representen un menor riesgo de pérdida de inversión para el gasto público. Así también, las entidades financieras privadas que capturen parte de este mercado puedan flexibilizar su confianza en estudiantes de menores ingresos, pero con habilidades académicas que potencian la protección de la inversión.

Abrazar un nuevo enfoque de predicción de default en créditos educativos, puede generar incentivos hacia todas los participantes del mercado de la educación superior, tanto demandantes como oferentes pueden aprovechar la información adicional como señal de un mejor resultado en sus intereses individuales que en conjunto benefician a la sociedad.

ANEXOS

Tabla A.1 Regresión logística

	Variable dependiente: default a 60 días			
género	-0.0311	-0.118	-0.122	-0.118
	(-1.14)	(-1.71)	(-1.75)	(-1.60)
edad	0.0217***	-0.0199	-0.0206	-0.0208
	(5.08)	(-1.38)	(-1.42)	(-1.43)
estrato 0	0.124	1.353	1.365	1.341
	(0.89)	(1.64)	(1.65)	(1.62)
estrato 1	-0.112	0.0674	0.0576	0.0517
	(-1.00)	(0.12)	(0.11)	(0.09)
estrato 2	-0.105	0.00781	0.00103	-0.00613
	(-0.96)	(0.01)	(0.00)	(-0.01)
estrato 3	-0.135	0.0161	0.0115	0.00436
	(-1.22)	(0.03)	(0.02)	(0.01)
estrato 4	0.0612	0.274	0.274	0.281
	(0.55)	(0.49)	(0.49)	(0.50)
Colegio privado	-0.0431	-0.0398	-0.0381	-0.0451
	(-1.38)	(-0.40)	(-0.38)	(-0.45)
Auto		0.0626	0.0638	0.0685
		(0.73)	(0.75)	(0.80)
Internet		0.00151	0.00312	0.000576
		(0.02)	(0.04)	(0.01)
Educación padre		-0.106	-0.108	-0.112
		(-1.53)	(-1.55)	(-1.60)
Educación madre		-0.0639	-0.0632	-0.0555
		(-0.81)	(-0.80)	(-0.70)
1-2 Hermanos		0.0871	0.0873	0.0861
		(0.81)	(0.81)	(0.80)
3-4 Hermanos		-0.0329	-0.0340	-0.0466
		(-0.26)	(-0.27)	(-0.37)
Puntaje icfes			-0.000502	-0.000283
			(-0.41)	(-0.23)
Universidad			0.0120	0.0417
			(0.07)	(0.25)
Salud				11.25
				(0.01)
Ciencias sociales				11.28
				(0.01)
Economía				11.40
				(0.02)
Ingeniería				11.28
				(0.01)
_cons	-2.051***	-1.459*	-1.305	-12.70
	(-14.89)	(-2.34)	(-1.73)	(-0.02)
N	44094	7494	7494	7474
t statistics in parentheses				
* p<0.05, ** p<0.01, *** p<0.001				

Elaboración propia

Los anteriores resultados pueden compararse con los encontrados en la regresión logística expuesta en la tabla A.1, donde se puede observar que ninguna característica a diferencia de la edad es significativa, sin

embargo, los signos son coherentes con lo encontrado en otros estudios, como que el asistir a un colegio privado reduce la probabilidad de incurrir en default en un préstamo educativo.

Aunque las variables no toman el peso estadístico que se esperaba dadas las limitaciones de la estimación logit en grandes muestras, es importante resaltar los resultados de las magnitudes del estrato socioeconómico, para los estratos 1, 2 y 3 la probabilidad de incurrir en default es menor que un estudiante de estrato 4. Esto puede explicarse por varias razones, una de ellas es que puede existir un sesgo de selección que no se aborda en este trabajo y que viene dado por el hecho de que los préstamos de ICETEX son enfocados especialmente en la población de estratos bajos.

El puntaje del examen Saber 11 toma el sigo esperado, ya que ante mayor sea el puntaje obtenido en las pruebas se reduce la probabilidad de incurrir en incumplimiento, esto confirma que otorgar préstamos a los estudiantes con altas habilidades independientemente de su nivel socioeconómico, representa un menor riesgo con respecto a otorgar créditos a estudiantes con bajas habilidades y con estratos socioeconómicos mayores.

Tabla A.2 Variables utilizadas

Variable	*Descripción*
Días de mora del crédito	Default a 30 ó 60 días
Nivel de sisben	0, 1, 2 o 3
Nivel de estrato	0, 1, 2, 3, 4, 5 o 6
Tipo de crédito	Tipo de crédito elegido por el estudiante
Educación de los padres	Niel educación alcanzada por los padres
Ocupación de los padres	Ocupación de los padres
Numero de hermanos	Numero de hermanos del estudiante
Tipo de colegio	Público/Privado
Valor pensión colegio	Valor pagado por pensión en colegio
Jornada del colegio	Completa, nocturna
Tipo de universidad	Público/Privado
Área del conocimiento	Área del conocimiento del programa
Puntaje global	Puntaje global de las pruebas Saber 11
Puntajes de lectura crítica, matemáticas e inglés	Pruebas Saber 11
Trabaja	Estudiante trabaja
Auto	Tiene automóvil
Internet	Tiene acceso a internet
Edad	Edad en el momento de solicitar el crédito
Género	Género del estudiante

Elaboración propia

REFERENCIAS

Attanasio, O., & Kaufmann, K. (2009). Educational choices, subjective expectations, and credit constraints (No. w15087). National Bureau of Economic Research.

Bagherpour, A. (2017). Predicting mortgage loan default with machine learning methods. University of California/Riverside.

Bardhan, P., & Udry, C. (1999). Development microeconomics. OUP Oxford.

Barone, S. (2006). Multivariate Analysis of Student Loan Defaulters at Prairie View A&M University. TG (Texas Guaranteed Student Loan Corporation).

Belfield, C. R. (2013). Student loans and repayment rates: The role of for-profit colleges. Research in Higher Education, 54(1), 1-29.

Bettinger, E., Gurantz, O., Kawano, L., Sacerdote, B., & Stevens, M. (2019). The Long-Run Impacts of Financial Aid: Evidence from California's Cal Grant. American Economic Journal: Economic Policy, 11(1), 64-94.

Bettinger, E., Gurantz, O., Kawano, L., Sacerdote, B., & Stevens, M. (2019). The Long-Run Impacts of Financial Aid: Evidence from California's Cal Grant. American Economic Journal: Economic Policy, 11(1), 64-94.

Carneiro, P., & Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. The Economic Journal, 112(482), 705-734.

Carneiro, P., & Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. The Economic Journal, 112(482), 705-734.

Carranza, J. E., & Ferreyra, M. M. (2019). Increasing higher education access: Supply, sorting, and outcomes in Colombia. Journal of Human Capital, 13(1), 95-136.

Chapman B. (2014) Income Contingent Loans: Background. In: Chapman B., Higgins T., Stiglitz J.E. (eds) Income Contingent Loans. International Economic Association Series. Palgrave Macmillan, London.

Chapman, B., & Doan, D. (2019). Introduction to the Special Issue "Higher Education Financing: Student Loans".

Chapman, B., & Doris, A. (2019). Modelling higher education financing reform for Ireland. Economics of Education Review, 71, 109-119.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Dearden, L. (2019). Evaluating and designing student loan systems: An overview of empirical approaches. Economics of Education Review, 71, 49-64.

Del Rey, E. (2001). Teaching versus research: a model of state university competition. Journal of Urban Economics, 49(2), 356-373.

Del Rey, E., & Verheyden, B. (2011). Loans, insurance and failures in the credit market for students.

Diris, R., & Ooghe, E. (2013). Financing Higher Education in Europe. Euroforum policy paper 8, 1-36.

Dynarski, M. (1991). Analysis of Factors Related to Default.

Ferreyra, M. M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., & Urzúa, S. (2017). Momento decisivo: La educación superior en América Latina y el Caribe.

Flint, T. A. (1997). Predicting student loan defaults. The Journal of Higher Education, 68(3), 322-354.

Friedman, M. (1955). The role of government in education.

García, V., Marqués, A. I., & Sánchez, J. S. (2012). Non-parametric statistical analysis of machine learning methods for credit scoring. In Management Intelligent Systems (pp. 263-272). Springer, Berlin, Heidelberg.

Goksu, A., & Goksu, G. G. (2015). A comparative analysis of higher education financing in different countries. Procedia Economics and Finance, 26, 1152-1158.

Goksu, A., & Goksu, G. G. (2015). A comparative analysis of higher education financing in different countries. Procedia Economics and Finance, 26, 1152-1158.

Greene, L. L. (1989). An economic analysis of student loan default. Educational Evaluation and Policy Analysis, 11(1), 61-68.

Gross, J. P., Cekic, O., Hossler, D., & Hillman, N. (2009). What Matters in Student Loan Default: A Review of the Research Literature. Journal of Student Financial Aid, 39(1), 19-29.

Herr, E., & Burt, L. (2005). Predicting Student Loan Default for the University of Texas at Austin. Journal of Student Financial Aid, 35(2), 27-49.

Herrera, S., & Pang, G. (2005). Qué tan eficiente es el gasto público en educación. Revista ESPE, 136-201.https://www.semana.com/educacion/articulo/es-hora-de-replantear-la-financiacion-de-las-universidades/623344

Kesterman, F. (2006). Student Borrowing in America: Metrics, Demographics, Default Aversion Strategies. Journal of Student Financial Aid, 36(1), 34-52.

Knapp, L. G., & Seaks, T. G. (1992). An analysis of the probability of default on federally guranteed student loans. The review of economics and statistics, 404-411.

Lleras, M. P. (2007). Investing in human capital: A capital markets approach to student funding. Cambridge University Press.

Lochner, L., & Monge-Naranjo, A. (2011). Credit constraints in education.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24-39.

Marx, B. M., & Turner, L. J. (2019). Student loan nudges: Experimental evidence on borrowing and educational attainment. American Economic Journal: Economic Policy, 11(2), 108-41.

Miller, K. (2017). Deeper in Debt: Women and Student Loans. American Association of University Women.

Mimoun, M. B. (2008). Credit constraints in education: Evidence from international data. Journal of Applied Economics, 11(1), 33-60.

Monteverde, K. (2000). Managing student loan default risk: Evidence from a privately guaranteed portfolio. Research in higher education, 41(3), 331-352.

Murphy, R., Scott-Clayton, J., & Wyness, G. (2019). The end of free college in England: Implications for enrolments, equity, and quality. Economics of Education Review, 71, 7-22.

Navarro, S. (2011). Using observed choices to infer agent's information: reconsidering the importance of borrowing constraints, uncertainty and preferences in college attendance (No. 2011-8). CIBC Working Paper.

Osunade, O., Ojo, O. M., & Ahisu, E. V. (2009). The role of internet on the academic performance of students in tertiary institutions. Journal of Educational research in Africa/Revue en Africanie de recherché en Education (JERA/RARE), 1(1), 30-35.

Palacios, M. (2002). Human capital contracts.

Palacios, M., DeSorrento, T., & Kelly, A. P. (2014). Investing in value, sharing risk: Financing higher education through income share agreements. AEI Paper & Studies.

Pizarro Milian, R., Zarifa, D., & Seward, B. Paying back student loans: Demographic, human capital and other correlates of default and repayment difficulty. Higher Education Quarterly.

S. Armstrong et al.Economics of Education Review(2019)

Salmi J. (2014) The Challenge of Sustaining Student Loans Systems: Lessons from Chile and Colombia. In: Chapman B., Higgins T., Stiglitz J.E. (eds) Income Contingent Loans. International Economic Association Series. Palgrave Macmillan, London

Schiefelbein, E. (1987). Education costs and financing policies in Latin America. World Bank, Education and Training Department, Operations Policy Staff.

Schiefelbein, E., & McGinn, N. F. (2017). Learning to educate: Proposals for the reconstruction of education in developing countries. Springer.

Scott-Clayton, J. (2012). Information constraints and financial aid policy (No. w17811). National Bureau of Economic Research. doi, 10, w17811.

Solis, A. (2011). Credit constraints for higher education.

Steiner, M., & Teszler, N. (2005). Multivariate Analysis of Student Loan Defaulters at Texas A&M University. TG (Texas Guaranteed Student Loan Corporation).

Steiner, M., & Tym, C. (2005). Multivariate Analysis of Student Loan Defaulters at the University of South Florida. TG (Texas Guaranteed Student Loan Corporation).

Volkwein, J. F., & Szelest, B. P. (1995). Individual and campus characteristics associated with student loan default. Research in higher education, 36(1), 41-72.

Wilms, W. W., Moore, R. W., & Bolus, R. E. (1987). Whose fault is default? A study of the impact of student characteristics and institutional practices on guaranteed student loan default rates in California. Educational Evaluation and Policy Analysis, 9(1), 41-54.

Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forests. TELKOMNIKA Indonesian Journal of Electrical Engineering, 10(6), 1519-1525.