Pronóstico del cumplimiento de pago de los clientes usando aprendizaje automático

Zuleyca Carolina Campos Cortesía

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Sistemas e Informática

Medellín, Colombia

2020

Pronóstico del cumplimiento de pago de los clientes usando aprendizaje automático

Zuleyca Carolina Campos Cortesía

Trabajo de grado presentado como requisito parcial para optar al título de:

Magíster en Ingeniería – Ingeniería Analítica

Director:

Prof. Juan David Velásquez Henao, PhD

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Sistemas e Informática

Medellín, Colombia

2020

A mi esposo Sergio Navarro,

Tu apoyo fue incondicional para concluir esta meta, estuviste siempre a mi lado incluso dando más de lo que podías dar.

A mis padres,

Por ustedes soy quien soy, todas mis metas alcanzadas reflejan el amor, confianza, educación que ustedes depositaron en mí.

Agradecimientos

A Dios por poner siempre oportunidades de éxito en mi camino y darme la valentía para tomarlas, poniéndome pruebas en cada situación, con el fin de enseñarme y crecer todos los días como persona. Gracias a él he culminado con éxito esta gran etapa en mi vida.

A Sergio Navarro por acompañarme hombro a hombro, sin ti, este camino no hubiera sido igual. Gracias por el apoyo que me brindaste desde el día que te dije “Quiero estudiar” e hicimos todo para que así fuera. Mi desvelos eran tus desvelos, mis viajes eran tus viajes, mis calificaciones las recibías como tuyas. Gracias por hacer este proyecto de vida tuyo.

Al profesor Juan David Velásquez, director de mi tesis y persona decidida a compartir sus conocimientos y experiencias. Gracias por acompañarme y guiarme en cada situación presentada a lo largo de la realización del proyecto.

A mi hermana Zulyber Campos, quien a pesar de estar a mi lado solo en el último año de la carrera, no dudó ni un segundo en brindarme su apoyo en lo que necesitara, dando siempre lo mejor de ella para mí.

A Ana María Giraldo y Marcela Gutiérrez, quienes me apoyaron desde que tome la decisión de estudiar, brindando ayudas en temas económicos y de flexibilidad de horarios, para poder llevar a cabo este proyecto, más que jefes han sido amigas en todo este camino.

A la Universidad Nacional de Colombia y al Estado Colombiano en general, por abrirle las puertas a una Venezolana, que llegó con todas las ganas de seguir creciendo personal y profesionalmente y aportarle a este país como lo hubiera hecho en el suyo.

A demás familiares, amigos y compañeros de trabajo, que pusieron su granito de arena para que este logro fuera posible.

Resumen

El score de cartera se presenta como un apoyo para la labor realizada por el área de cartera de la empresa Créditos Orbe. Con el score de cartera, se busca dirigir los recursos y las estrategias de cobro a grupos de clientes determinados por su probabilidad de pago, lo cual impacta positivamente en la eficacia del recaudo y la eficiencia del proceso. La revisión de literatura expone que solo el 1,5% de los modelos de score en los últimos 10 años están enfocados en la cobranza; estos modelos son usados generalmente para la colocación del crédito. En este trabajo se realiza un modelo para score de cartera el cual es entrenado a partir de la información histórica de comportamiento de pago de los clientes con créditos activos en la empresa. La empresa actualmente cuenta con un modelo de regresión logística, el cual usan para dirigir estrategias de cobranza. Se plantean modelos alternativos que pueden mejorar la precisión del modelo actual, a la luz de los hallazgos de la revisión de literatura. Finalmente, después de entrenar modelos de vecino más cercano, regresión logística; y modelos de ensamble como bosques aleatorios, voto mayoritario, adaboost y bagged con árboles de decisión; el prototipo basado en bosques aleatorios es el que mejor se desempeña, superando las métricas del modelo actual. Se le hace entrega a la empresa del prototipo para su uso semanal, así como también, de las herramientas necesarias para monitorearlo y generar alertas de rendimiento.

Palabras clave: Clasificación, score de cartera, aprendizaje automático, bosques aleatorios, gestión de cobranza.

Forecasting customer payment compliance using machine learning

Abstract

The collection score is presented as a support for the work carried out by the collection area of Créditos Orbe company. The collection score seeks to direct collection resources and strategies to groups of clients determined by their probability of payment, which positively impacts the effectiveness of collection and the efficiency of the process. The literature review shows that only 1.5% of the score models in the last 10 years are focused on collection; These models are generally used for credit placement. In this work, a collection score model is carried out, which is trained based on the historical information of the payment behavior of clients with active credits in the company. The company currently has a logistic regression model, which they use to direct collection strategies. Alternative models are proposed that can improve the precision of the current model, in light of the findings of the literature review. Finally, after training K-nearest neighbor models, logistic regression; and assembly models such as random forests, majority vote, adaboost and bagged with decision trees; the prototype based on random forests performs best, exceeding the metrics of the current model. The prototype is delivered to the company for weekly use, as well as the necessary tools to monitor it and generate performance alerts.

Keywords: Classification, collection score, machine learning, random forest, collection management

Contenido

Resumen ..... **ix**

Contenido ..... **xii**

Lista de tablas ..... **xiv**

Lista de figuras ..... **xv**

1. Introducción **1**

2. Definición del problema de negocio **3**

3. Definición del problema de analítica **9**

4. Información Disponible y comprensión de datos **11**

4.1 Información disponible 11

4.2 Compresión de datos 15

5. Revisión sistemática de literatura **23**

5.1 Metodología 24

5.1.1 Proceso de búsqueda 24

5.1.2 Criterios de inclusión y exclusión 24

5.1.3 Información recopilada 25

5.2 Resultados de la búsqueda 26

5.3 Discusión 28

5.3.1 Q1: ¿Cuáles son las técnicas de aprendizaje estadístico o aprendizaje automático más exitosas aplicadas a la gestión de cobranza? 28

5.3.2 Q2: ¿Qué porcentaje de los artículos recuperados están enfocados al área de cobranza? 29

5.3.3 Q3: ¿Qué variables se ha usado en los estudios realizados? 29

5.3.4 Q4: ¿En qué tipos de negocio se han aplicado estas técnicas? 30

5.3.5 Q5: ¿Cuáles son los hallazgos recientes basados en el uso de aprendizaje estadístico o aprendizaje automático sobre la calificación crediticia? 30

6. Justificación de este trabajo **33**

7. Hipótesis 35

8. Objetivos de investigación 37

8.1 Objetivo General 37

8.2 Objetivos Específicos 37

9. Inventario de modelos disponibles 39

9.1 Vecino más cercano 39

9.2 Regresión logística 39

9.3 Bosques aleatorios 39

9.4 Voto mayoritario 40

9.5 Adaboost 40

9.6 Bagged con árboles de decisión 40

10. Calibración y validación de modelos 43

11. Recalibración del modelo 47

12. Evaluación del negocio 49

13. Distribución 55

14. Conclusiones 57

14.1 Objetivos 57

14.1.1 Objetivo 1: Evaluar si existen modelos alternativos que sean más precisos a la luz de las métricas de calidad 57

14.1.2 Objetivo 2: Establecer parámetros que permitan determinar cuándo se deben autocalibrar los modelos. 57

14.1.3 Objetivo 3: Implementar un prototipo con el nuevo modelo estimado basado en el sistema autocalibrado. 58

14.2 Evaluación de la Hipótesis 58

15. Referencias 61

Lista de tablas

Tabla 1. Variables disponibles para la construcción de modelos. 12

Tabla 2. Descripción de los artículos seleccionados..... 26

Tabla 3. Resultados de las corridas de los modelos. 44

Tabla 4. Efectividades de las estrategias propuestas. 52

Lista de figuras

Figura 1. Indicador de Morosidad de la empresa Créditos Orbe. 5

Figura 2. Variables con valores nulos. 15

Figura 3. Distribución de la variable respuesta..... 16

Figura 4. Distribución de la cartera por días en mora..... 17

Figura 5. Distribución de la cartera por clientes ilocalizados. 17

Figura 6. Distribución de clientes ilocalizados según la zona 18

Figura 7. Distribución de estado civil según valor de la cuota promedio que paga el cliente. 18

Figura 8. Distribución de los clientes según número de veces que no llegan a un acuerdo con el agente del call center..... 19

Figura 9. Distribución de los clientes según número de veces que ha necesitado una visita de cobro. 20

Figura 10. Distribución de los clientes según estado del último acuerdo de pago……..21

Figura 11. Distribución de los créditos según si tiene un codeudor asociado. 21

Figura 12. Ranking de los motivos de mora más frecuentes. 22

Figura 13. Distribución de la probabilidad de pago de los clientes de Créditos Orbe con los resultados del prototipo. 49

Figura 14. Distribución de la probabilidad de pago de los clientes de Créditos Orbe con los resultados del modelo actual de la empresa..... 50

Figura 15. Estrategias de cobro por cada segmento según el prototipo. 51

Figura 16. Libro de Jupyter para correr el prototipo..... 55

Figura 17. Libro de Jupyter para evaluar las métricas del modelo..... 56

1. Introducción

La realidad de la economía colombiana radica en que las entidades financieras, en su mayoría, solo prestan a personas que ya llevan una vida crediticia construida, quedando aquellos clientes, como, empleados informales, microempresarios, entre otros, por fuera de estas clasificaciones y con mayor dificultad a la hora de acceder a un préstamo.

Créditos Orbe SAS es una empresa que otorga créditos para motocicletas, que nació de la necesidad de un segmento de la población que quería adquirir su vehículo, y las entidades no le otorgaban dichos créditos. El cliente objetivo de la empresa, son aquellas personas no bancarizadas, que no tienen los soportes para demostrar sus ingresos porque son independientes, informales, o microempresarios.

Por lo anterior, la empresa es un poco más flexible en cuanto a las políticas de crédito que evalúa en el otorgamiento de los mismos, logrando llegar a este segmento de clientes; sin embargo, esta premisa arrastra dos problemáticas, y es que este cliente, es más vulnerable a que a lo largo de la vida del crédito, no pueda cumplir con el pago de su cuota; y la segunda es que por dicha flexibilidad, no solo llega a este segmento de clientes, sino que también, a un grupo de clientes que solo quieren la facilidad del proceso para obtener el vehículo y después la recuperación de esta cartera es un trabajo arduo.

Hasta el segundo semestre del año 2018, la empresa no había logrado ser eficiente en su gestión de cobranza para mantener los niveles adecuados de morosidad, por lo que, con el fin de mejorar tal gestión y de optimizar los recursos de la empresa, se implementó un score de cobranza, basado en un modelo de regresión logística. Sin embargo, dado los resultados que se han obtenido, la empresa plantea la necesidad de mejorar dicho modelo y poder utilizar el score de cartera, de manera oportuna en el día a día.

Tener este tipo de herramientas, apuntan a la eficacia del recaudo, pues las estrategias de cobro no serían de forma masiva, es decir, que todos los agentes del call center gestionen la totalidad de la cartera de la empresa a través de llamadas, lo que genera un alto costo y un uso totalmente ineficiente de los recursos; el agente del call center no invertiría mucho tiempo en hacer la revisión de toda la información de los clientes, la cual muchas veces no es oportuna, siendo su foco y prioridad, la generación de recaudos, que debe ser el objetivo principal de un call center.

En el presente trabajo, se exploran modelos alternativos a la regresión logística, que permitan cuantificar, de mejor forma, la probabilidad de que un cliente en mora cumpla con su obligación, con base a la comparación de las métricas de calidad del modelo actual.

El presente documento está estructurado de la siguiente manera: los Capítulos 1 y 2, hacen referencia a la definición del problema de negocio y del problema de analítica respectivamente; el Capítulo 3 presenta la información disponible para el entrenamiento del modelo y una breve compresión de los mismos. En el Capítulo 4 se presenta la revisión sistemática de literatura, la metodología usada y los resultados obtenidos; El Capítulo 5 presenta la justificación del trabajo. Los Capítulos 6 y 7, hacen referencia a la hipótesis y los objetivos de investigación; En el Capítulo 8 se muestra un inventario de modelos disponibles; el Capítulo 9 presenta la calibración y validación de los modelos; en el Capítulo 10 se explica cómo será la recalibración del modelo. Los Capítulos 11 y 12 presentan la evaluación del modelo en términos del negocio y la distribución del prototipo, respectivamente; para finalizar en el Capítulo 13 con las conclusiones del trabajo.

2. Definición del problema de negocio

Las empresas son unidades económicas que desarrollan sus actividades con el fin de alcanzar sus objetivos estratégicos previamente planificados. Estas actividades deben realizarse en un marco de equilibrio tanto económico como financiero. Todas las empresas que otorgan créditos están expuestas al riesgo crediticio, que según Saavedra García (2010) es la probabilidad de que el prestatario incumpla con sus obligaciones. Esto impacta directamente sobre la cartera que es el recurso sobre el cual gira la liquidez de las mismas, por lo que, la adecuada administración de este riesgo, es lo que les permite a las empresas alcanzar rentabilidad y sostenibilidad.

La morosidad es el incumplimiento de los plazos legales de pago, lo que lleva a pagos vencidos. Es un tema de especial relevancia para las empresas, principalmente por el hecho de que puede provocar la insolvencia de quien la padece, y en consecuencia, supone un desajuste en los pagos previstos que deben hacer frente al acreedor. El índice de morosidad se define como la proporción de cartera vencida respecto a la cartera total de la entidad (Chavarín Rodríguez, 2015).

Colombia no escapa de la morosidad; según la Federación Latinoamericana de Bancos (FELABAN), con corte a septiembre de 2018, Colombia tiene el indicador más alto en calidad de cartera, en comparación con los demás países de la región, con un valor de 4,91%, el cual ha venido creciendo, al compararlo con el corte de septiembre de 2017 el cual era 4,37% y en diciembre de 2016 cerró en un 3,1%.

En un estudio de TransUnion, según cifras del DANE, el desempleo en áreas metropolitanas ha aumentado en los últimos años, lo que podría afectar la capacidad de pago de los clientes. También, según información del Banco de la República, el crecimiento anual de los salarios reales se desaceleró, volviéndose negativo en el 2018. Por otra parte, las tasas de interés han venido bajando lo que traduce en disminución del costo de los créditos y las exigencias en las políticas de créditos se

han vuelto menos restrictivas (Olivella, 2018). Según cifras de TransUnion en el segundo semestre de 2018:

- Los indicadores de mora muestran un deterioro en las principales líneas de negocio, las líneas que tuvieron un deterioro mayor son microcrédito 11,7% (+113pbs) y vehículo 5,6% (+86pbs).
- Los niveles de deuda promedio continuaron creciendo indicando que los consumidores continúan usando crédito.
- Para el sector de vehículo, la participación de consumidores de alto riesgo ha aumentado en los últimos años.
- Las originaciones de créditos de vehículos crecieron 7%.

El riesgo evidenciado anteriormente, también es notorio en la empresa Créditos Orbe SAS. El indicador de cartera de vencida es el saldo de los créditos que presentan mora mayor a 30 días sobre el saldo total de la cartera. El valor de dicho indicador es presentado en la Figura 1, y a cierre del 2019 es de 12,65%. Este indicador se evalúa junto con el indicador de cartera reciente, que corresponde a los créditos que llevan menos de 6 meses de colocados; y el indicador de cartera madura que son los créditos que tienen más de 6 meses de colocados, dichos indicadores a cierre de 2019 fueron 17,38%% y 5,78% respectivamente.

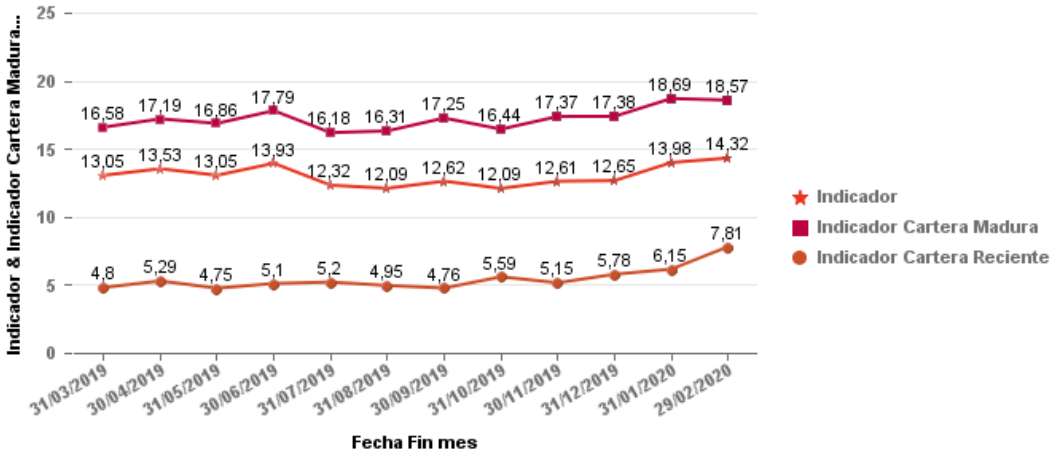

Figura 1. Indicador de Morosidad de la empresa Créditos Orbe.

El indicador de cartera castigada es uno de los más importante para la empresa, ya que, la superfinanciera y las entidades que dan el fondeo, monitorean dicho indicador; se mide saldo de los créditos que superan los 120 días en mora sobre el saldo total colocado; Créditos Orbe cerró el 2019 con un indicador del 9,49%, el cual ha seguido una tendencia creciente a lo largo del tiempo, y durante el año 2019, se ha logrado mantener, pero no se ha logrado disminuir, a pesar de todas las estrategias de cobro realizadas.

La cobranza es un servicio importante que permite el mantenimiento de los clientes al igual que da la posibilidad de volver a prestar; es un proceso estratégico y clave para generar el hábito y una cultura de pago en los clientes. Hay que diferenciar las estrategias de cobro para afrontar la morosidad; los cobradores deben estar atentos en la forma de abordar al cliente, ya que un estilo de comunicación incorrecto puede dañar la relación con los mismos. Se debe comprender al cliente y las razones por la cual no paga para poder elegir la acción adecuada, presionar demasiado cuando es posible que solo tenga problemas de flujo de efectivo a corto plazo podría causar la pérdida de un cliente, además de que se invierte mucho tiempo y dinero en esta

operación Onar, Oztaysi & Kahraman (2018). Por ello, es necesario establecer diferencias entre distintos tipos de clientes y definir cuidadosamente las estrategias a seguir con cada uno de ellos. Esto no se puede afrontar de manera aislada, una de las principales maneras de mitigar los riesgos de impago y de acompañar las estrategias de cobro, es la gestión de la cartera. Las empresas en la actualidad buscan ser cada vez más eficientes en dicha gestión.

Wittlinger, Carranza & Mori (2008) define gestión de cobranza como el "conjunto de acciones coordinadas y aplicadas adecuada y oportunamente a los clientes para lograr la recuperación de los créditos, de manera que los activos exigibles de la institución se conviertan en activos líquidos de la manera más rápida y eficiente posible, manteniendo en el proceso la buena disposición de los clientes para futuras negociaciones."

La gestión de cobranza busca recuperar a tiempo la totalidad de la cartera, usando diferentes herramientas, enfocadas, oportunas y personalizadas según las características de los clientes, es por esto, que puede variar las acciones a tomar, como, el tipo de mensaje que se debe usar tanto para las comunicaciones escritas como las llamadas, qué tono de voz se debe usar en las llamadas telefónicas y a qué hora del día se debe llamar, la experticia del agente, entre otros factores.

Para Hector (2011) otro factor a tener en cuenta es la selección del tipo de comunicación según las características del cliente, pues afecta directamente el costo de la operación; por ejemplo, mensajes de texto y correos electrónicos se pueden enviar masivamente a un bajo costo, mientras que una llamada tiene el costo del proveedor de servicio telefónico más el asesor que debe manejar la llamada. Todas estas acciones que se pueden implementar deben ir en un marco de planeación y estrategias definidas del proceso de cobranza.

Algunas de las estrategias de las empresas se enfocan en la optimización de esfuerzos y recursos de cobranza aplicando diversos modelos estadísticos, matemáticos, o de ingeniería de datos, que le permitan mejorar el desempeño de la

cartera, mediante la administración de los límites de una cuenta, establecer cobros basados en el riesgo y estrategias de recuperación, retener clientes estables, predecir los clientes que definitivamente no pagaran, ayudar a mejorar la orientación de las ofertas y para gestionar las tasas de recuperación (Sarlija, Bensic & Zekic-Susac, 2009).

Las organizaciones necesitan sistemas efectivos de cobro de deudas, un sistema exitoso requiere del uso de dichos modelos que permiten tomar decisiones de forma rápida y automatizada, además de posibilitar el desarrollo de diferentes estrategias para asegurar el pago. Si se logra predecir con precisión los clientes retrasados de los clientes realmente morosos, el departamento de cobranza solo contactara intensivamente a este segmento de clientes (Sun, Wiering & Petkov, 2014).

El score de cartera, se presenta como una oportunidad de mejora para la labor en la gestión de cobro, pues busca enfocar las estrategias persuasivas de cobro a nichos de cartera caracterizados por riesgo de pago, lo cual incide positivamente en la eficacia del recaudo.

Dichos modelos proponen automatizar el proceso de gestión de cobro en cuanto a usar o no una determinada operación y estrategia, reduciendo ineficiencias asociadas a subjetividad, esto, dependiente a un conjunto de variables relevantes de decisión. El éxito de este modelo está en la calidad del algoritmo utilizado y de la existencia de un sistema eficiente de análisis de datos.

La empresa Créditos Orbe cuenta con 45 agentes de call center, los cuales tienen una capacidad para gestionar 55 clientes por día, lo que equivale a 2.475 clientes gestionados al día, por lo que en un mes un cliente solo tendría 2 gestiones de cobro. Con el uso de modelos que permitan segmentar esta cartera, los agentes no tendrían que llamar a toda la cartera, sino solo a los clientes que realmente ameriten una gestión de cobro, pues habrá un porcentaje que se podrá gestionar de forma automática a través de mensajes, correos y robots.

Dado lo anterior, el objetivo es clasificar a los clientes según su probabilidad de incumplimiento de pago, obteniendo segmentos de clientes a los cuales se les aplicará una estrategia de cobro particular. Este proyecto le permite a la empresa ser eficiente con las estrategias de cartera a utilizar y los recursos que las implementan, reduciendo costos y convirtiendo pérdidas en ingresos.

Se establecen como criterios de éxito la posibilidad de tener segmentos de clientes, para lograr ser eficientes en las herramientas y estrategias de cobro a utilizar; y mejorar el ajuste del modelo actual, obteniendo una mejor discriminación de los clientes.

3. Definición del problema de analítica

La empresa créditos Orbe SAS ha diseñado un score de cartera mediante un modelo de regresión logística, lo que le permite clasificar a los clientes morosos, logrando establecer adecuadamente las prioridades de cobro, enfocando los recursos en los clientes con riesgo medio y bajo, y usando herramientas de marcación automática en clientes con riesgo bajo y alto. A pesar de tener buenos resultados con los clientes de riesgo bajo, de los cuales se tiene recaudo después de marcación automática o mensajes de texto, con los clientes de riesgo medio, no se ha logrado la efectividad deseada, solo un 35%, y en esta clasificación se encuentra la mayor concentración de la cartera de la empresa.

A pesar de que el ajuste del modelo es del 71%, no es potente discriminando entre los clientes de riesgo medio o alto que representan el porcentaje más alto de la cartera, y el equipo actual de cobradores de la empresa, no podría realizar el número de gestiones necesarias para un recaudo, en este gran número de clientes. Dado que la totalidad de los clientes tienen comportamientos muy variados, por ejemplo, los que no se han podido contactar, los que no han pagado la primera cuota, los que pagan con un asesor presencial, los clientes recién desembolsados, entre otros, se tiene la hipótesis de que segmentando muy bien estos grupos y aplicando los modelos de aprendizaje automático o estadístico, se pueda tener una mayor discriminación de los mismos, apuntando a la eficiencia del proceso.

Adicionalmente, la empresa no ha logrado que el score de cartera funcione en tiempo real, este se actualiza una vez al mes, y la clasificación de los clientes se mantiene estática a lo largo del mismo, si el cliente va cambiando su comportamiento a lo largo del mes, no se ve reflejado hasta el siguiente, por lo que ese cliente recibe un tratamiento que no era el más adecuado según sus características.

Por los motivos descritos anteriormente, un proceso de análisis predictivo, sobre el comportamiento de clientes que han solicitado créditos, utilizando la información de los mismos y considerando también variables exógenas, apoya a la empresa a tomar decisiones más adecuadas acerca de la estrategia de cobro a utilizar según la segmentación de los clientes por su nivel de riesgo. Un mejor modelo de predicción, puede ayudar a reducir la exposición al riesgo de incumplimiento de pagos, e incrementar la rentabilidad de la cartera de la empresa.

El análisis estadístico y el aprendizaje automático, buscan y describen patrones estructurados, por lo que son herramientas que sirven para ayudar a explicar ciertos comportamientos de interés. Es por ello, que esta investigación se basa en explorar dichos modelos, los cuales permitan tener resultados de manera oportuna y mejorar la segmentación de clientes que existe hoy en la empresa, obteniendo así, usar los recursos de manera eficiente con las estrategias adecuadas.

En cuanto a recursos de software se dispone de SQLdeveloper para realizar tareas de extracción sobre una base de datos Oracle 11g. Para realizar las tareas de minería de datos se cuenta con softwares libres. Las técnicas que se van a emplear son de analítica predictiva.

4. Información Disponible y comprensión de datos

4.1 Información disponible

Los datos utilizados corresponden a los clientes de la empresa Créditos Orbe SAS, a los que se les otorgó crédito desde el año 2015, datos que van desde los sociodemográficos del cliente, de buró de crédito y de habito de pago en la obligación activa. Como restricción, los datos son confidenciales y la empresa los pone a disposición firmando una cláusula de confidencialidad, en donde queda expreso que la universidad puede publicar resultados, más no los datos propios de los clientes. La obtención de estos datos no supone ningún costo adicional ni para la empresa ni la universidad.

La cartera de la empresa está conformada por 52.748 clientes, de los cuales 32.086 aplican para gestiones de cobro internas, luego del procesamiento quedaron 27.395 clientes, de los cuales el 80% de los datos se usará para elaborar el modelo y el 20% restantes para su evaluación. Estos datos cubren los casos requeridos para la obtención de los resultados y no contienen errores dado que vienen directamente de una copia de la base de datos.

La data se construye a partir de 3 grandes tablas que son, la que contiene todas las variables sociodemográficas del cliente, la del comportamiento de pago de cliente y la de las gestiones donde está el registro de los acuerdos de pago del cliente.

A continuación, se listan las variables disponibles para la realización del trabajo:

Tabla 1. Variables disponibles para la construcción de modelos.

Variable	Descripción
Nro_Obligación	ID de los créditos
Zona	Zona en la cual de desembolsó el crédito
IndicadorIlocalizados	Identificador si el cliente está ilocalizado.
ValorMora	Saldo en Mora del crédito
NVIG	Número de cuotas que ha estado vigente el crédito
N30	Número de veces que ha estado en mora 30.
N60	Número de veces que ha estado en mora 60.
N90	Número de veces que ha estado en mora 90.
N120	Número de veces que ha estado en mora 120.
Mayor120	Número de veces que ha estado en mora mayo120.
ValorDesembolso	Valor por el cual al cliente se le realizó el desembolso del crédito
ValorUltimoPago	Valor que corresponde al último pago que realizó el cliente
ValorSaldoCapital	Valor que corresponde al saldo capital del crédito
ValorVehiculo	Valor del vehículo al momento de desembolsar el crédito
ValorCuota	Valor de la cuota del crédito mensual
Indicador Madurez	Indica cual es el mes de madurez que tiene el crédito
IndicadorRegistroGM	Indica si el crédito tiene registro de Garantía Mobiliaria
Madurez	Indica si la etapa en la que se encuentra el crédito
IndicadorEjecuciónGM	Indica si el crédito tiene Ejecución de Garantía Mobiliaria
IndicadorTieneAcuerdo	Indica si el cliente tiene un acuerdo de pago vigente
IndicadorFallecido	Indica si el cliente ha sido reportado como fallecido
IndicadorCodsinllamada	Indica si el crédito tiene un codeudor y no ha sido llamado
IndicadorSinllamada	Indica si al cliente no se le ha realizado llamada desde que cayó en mora
IndicadorRefSinllamada	Indica si a las referencias del crédito no se le ha realizado llamada
IndicadorNoSolVisita	Indica si al cliente no se le ha solicitado una visita de cobro
IndicadorLlamadaHorarios	Indica si al cliente se le ha llamado en horarios diferentes
IndicadorSolWhatsApp	Indica si al cliente se le ha enviado WhatsApp desde que cayó en mora
IndicadorCantidadVisitas	Número de veces que al cliente se le ha realizado visita de cobro
IndicadorTodosTelefonos	Indica si al cliente se le ha llamado a todos los teléfonos activos
IndicadorTodasDirecciones	Indica si al cliente se le ha visitado a todas las direcciones.

Tabla 1. (Continuación)

Variable	Descripción
MotivoMoraFreq	Indica el motivo de mora que con mayor frecuencia indica el cliente
IndicadorVisitadiferentedia	Indica si al cliente se le ha realizado visita de cobro en días diferentes
CantidadRenuente	Número de veces que no se ha logrado un acuerdo con el cliente
CantidadWhastApp	Número de veces que al cliente se le ha enviado mensajes de WhatsApp
IndicadorFacturaRecibida	Indica si el cliente le ha llegado la factura.
IndicadorLlamadaBienvenida	Indica si al cliente se le realizó la llamada de bienvenida
IndicadorSMSBienvenida	Indica si al cliente se le envío mensaje de bienvenida
IndicadorCeroPagos	Indica si el crédito no tiene ninguna cuota paga
IndicadorMotoRetenida	Indica si al cliente ya se le retuvo la moto
IndicadorUltAcuerdoPago	Indica el estado del último acuerdo de pago del cliente
CantidadAcuerdoCumplidos	Número de acuerdos que el cliente realiza y cu
CantidadAcuerdoIncumplidos	Número de acuerdos que el cliente realiza y no cumple
CantidadAcuerdosParcial	Número de acuerdos que el cliente cumple de forma parcial
IndicadorComplGestionado	Indica si el cliente ha sido gestionado de forma exitosa
IndicadorCercanoEjecutar	Indica cuando el cliente está próximo a ser castigado
IndicadorCasaCobro	Indica si el cliente fue asignado a una casa de cobro
DiasMora	Cantidad de días en mora que presenta el crédito
CuotasPagadas	Número de cuotas que ha pagado el cliente
CuotasSinPagar	Número de cuotas que le faltan al cliente para cancelar el crédito
Cuotas en Mora	Número de cuotas que el cliente tiene en mora
UltimoMotivoMora	Indica el último motivo de mora informado por el cliente
CantidadContactoNoExitoso	Número de veces que se ha llamado al cliente y no ha respondido
MesUltimoPago	Indica el último mes que el cliente hizo algún pago
MesesMora	Indica desde que mes el cliente ha estado en mora
Y	Indica si el cliente ha realizado pagos consecutivos en los últimos 3 meses
Multas	valor de multas del cliente al momento de solicitar el crédito
Edad	Edad del cliente al momento de solicitar el crédito
Genero	Género del cliente

Tabla 1. (Continuación)

Variable	Descripción
Creditos_vigentes	número de créditos vigentes que tenía el cliente según el reporte de centrales de riesgo al momento de la colocación
Estado_civil	Estado civil del cliente
Nivel_estudio	nivel de estudio del cliente
Tipo_vivienda	Tipo de vivienda del cliente
Personas_cargo	número de personas a cargo que tiene el cliente
Plazo	plazo del crédito
Calificación Asignación	Calificación interna de la empresa según el nivel de riesgo del cliente
Codeudor	si el crédito tiene o no codeudor
Ingresos	ingresos del cliente al momento de solicitar el crédito
Egresos	egresos del cliente al momento de solicitar el crédito
Subtipo_cliente	subtipo de cliente, si es formal o informal
Tipo_cliente	tipo de cliente, si es empleado, independiente, pensionado
Habito_pago	habito de pago del cliente de acuerdo a los informes de centrales de riesgo
Marca	marca de la moto que seleccionó el cliente
Referencia	referencia de la moto que seleccionó el cliente
Endeudamiento	saldo de endeudamiento del cliente según información de las centrales de riesgo al momento de la colocación
Calificación	calificación otorgada por las centrales de riesgo al cliente al momento de la colocación
Acierta	score que le otorgan las centrales de riesgo al cliente y tenía el cliente al momento de la colocación
Monitor	Calificación de centrales de riesgo de comportamiento de pago del cliente a la empresa vs sus otras obligaciones en otras entidades
Antigüedad	antigüedad laboral del cliente al momento de la colocación
Porc_inic	porcentaje de cuota inicial que da el cliente al momento del otorgamiento del crédito

Con ellos se pretende dar los interesados un modelo de clasificación de los clientes de la empresa, usando modelos de aprendizaje estadístico y automático, previamente vistos en trabajos similares.

4.2 Compresión de datos

Una vez descrita la información disponible, se procede a explorar los datos, lo que permite hacer un primer acercamiento, y determinar la consistencia y completitud de la información. La data estaba compuesta originalmente por 96 variables y 30.828 registros. La Figura 2 muestra que 18 de ellas tienen más del 5% de valores nulos, por lo que se procede a eliminarlas quedando 78 variables disponibles. A nivel de registros, se eliminan los que tienen valores nulos y valores atípicos validados por los expertos del área, quedando un total de 29.387 registros.

	Total	Percent
'REEX'_FREQ	29430	0.997627
'INPE'_FREQ	29374	0.995729
'ENFE'_FREQ	28657	0.971424
'INPA'_FREQ	28547	0.967695
'VIAJ'_FREQ	27980	0.948475
'SIVO'_FREQ	27781	0.941729
'CRET'_FREQ	27493	0.931966
'CADO'_FREQ	27016	0.915797
'OTGA'_FREQ	26241	0.889525
'DESE'_FREQ	25880	0.877288
'PANO'_FREQ	21512	0.729220
Nivel Estudio	19714	0.668271
'DESC'_FREQ	18425	0.624576
TIPO_CONTRATO	16218	0.549763
'ININ'_FREQ	14859	0.503695
SCORE_SIMU	8387	0.284305
Cantidad Personas Cargo	3454	0.117085
PROPOSITO_VEHICULO	1982	0.067186

Figura 2. Variables con valores nulos.

La variable respuesta corresponde a si el cliente ha pagado consecutivamente en los últimos 3 meses, es indispensable validar como es la distribución de la variable respuesta, para entender si la metodología debe considerar un balanceo de muestras, el problema de clasificación presentado es binario, y en la Figura 3, se puede evidenciar que no se tiene el problema de muestras desbalanceadas.

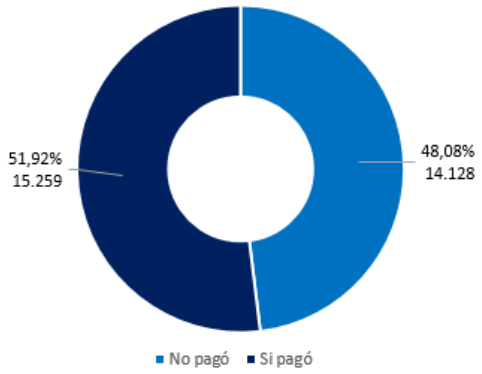

Figura 3. Distribución de la variable respuesta.

El 70,87% de la cartera tiene de 0 a 15 días en mora, el 12,60% está de 16 a 30 días en mora, el 10,91% tiene mora de 30 a 60 días y el 5,62% restante tiene mayor a 60 días en mora (Figura 4). La compresión de los datos que sigue a continuación, está enfocada en conocer como es el comportamiento de la cartera de la empresa. El 7% de los clientes están ilocalizados (Figura 5).

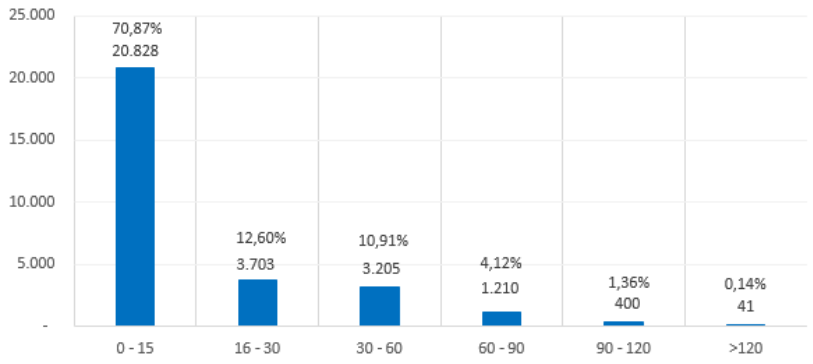

Figura 4. Distribución de la cartera por días en mora.

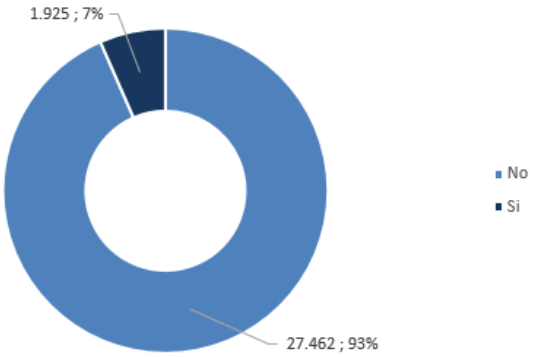

Figura 5. Distribución de la cartera por clientes ilocalizados.

La Figura 6 muestra el porcentaje de clientes ilocalizados por zona, Eje Cafetero es la zona más pequeña en cuanto a créditos activos, sin embargo es la que tiene el porcentaje más alto de clientes ilocalizados, seguida de la zona Centro del país. Bogotá es la zona con más créditos activos y menos clientes ilocalizados.

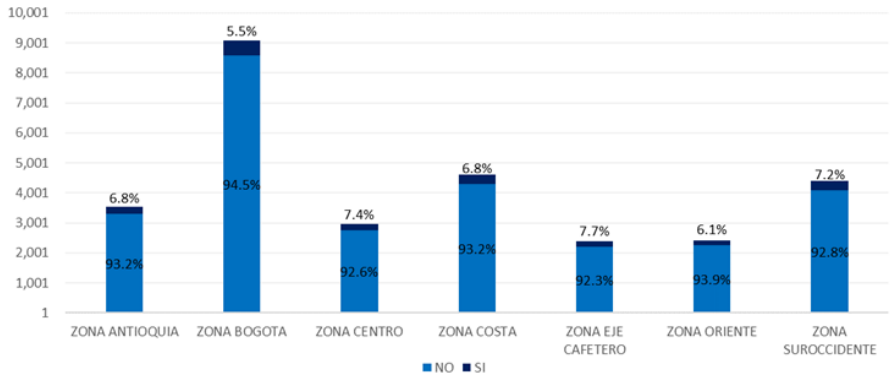

Figura 6. Distribución de clientes ilocalizados según la zona

La Figura 7 presenta que las personas de estado civil soltero pagan una cuota promedio mensual más alta, seguido de los clientes casados y unión libre. Los clientes de estado civil viudo son los que pagan una cuota promedio mensual más baja.

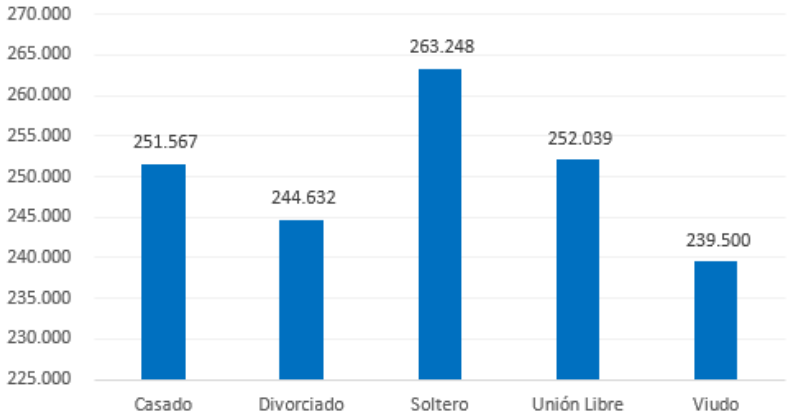

Figura 7. Distribución de estado civil según valor de la cuota promedio que paga el cliente.

Con el 3% de la cartera no se llegó a un acuerdo una vez, con el 1% dos veces, y con el 2% de la cartera al menos tres veces no se llegó a un acuerdo con el cliente. El 94% de la cartera no ha presentado esta dificultad (Figura 8)

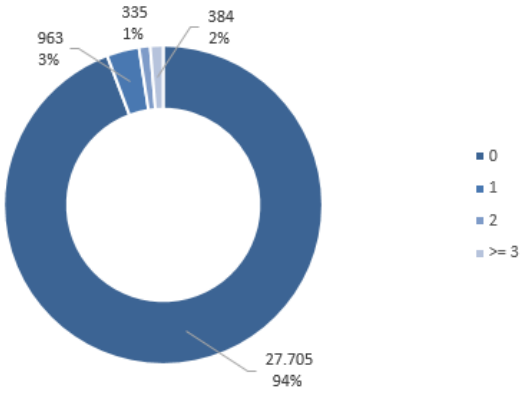

Figura 8. Distribución de los clientes según número de veces que no llegan a un acuerdo con el agente del call center.

La Figura 9 muestra la distribución de los clientes que han necesitado una visita de cobro, porque la gestión telefónica no ha sido suficiente, el 91% de los clientes, no ha necesitado una visita, el 4% de los clientes ha necesitado entre 1 y 2 visitas, el 3% entre 3 y 6 visitas y el 2% más de 7 visitas.

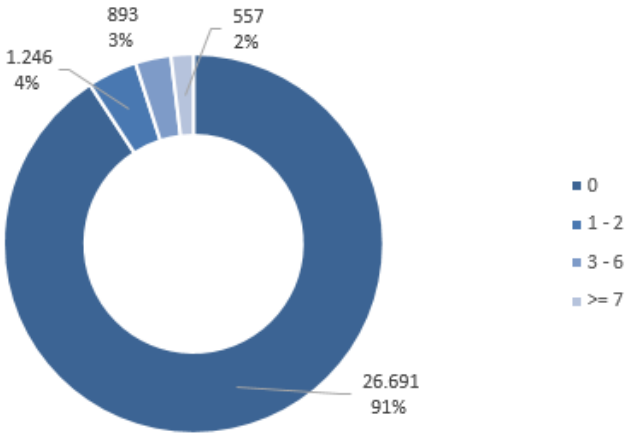

Figura 9. Distribución de los clientes según número de veces que ha necesitado una visita de cobro.

El 40% de los clientes no necesitan realizar un acuerdo de pago para realizar los pagos, por lo general que pagan antes de tener la necesidad de generar un acuerdo con ellos. El 32% de los clientes cumplieron el último acuerdo de pago que realizaron, el 24% de los clientes incumplieron el último acuerdo de pago realizado, y el 4% restante cumplió parcialmente el acuerdo (Figura 10). La Figura 11 indica que solo el 11% de los créditos de la empresa, tienen un codeudor asociado que respalde la deuda.

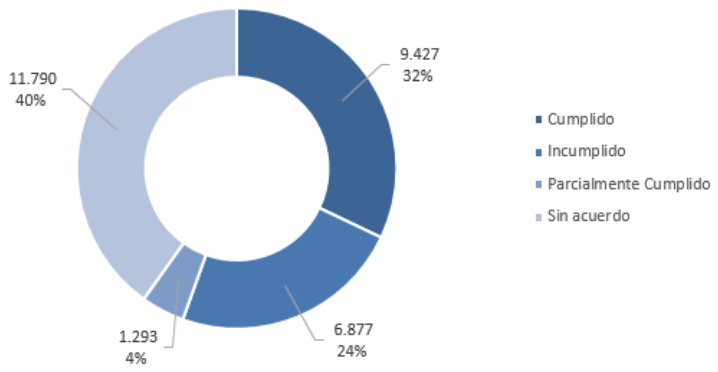

Figura 10. Distribución de los clientes según estado del último acuerdo de pago.

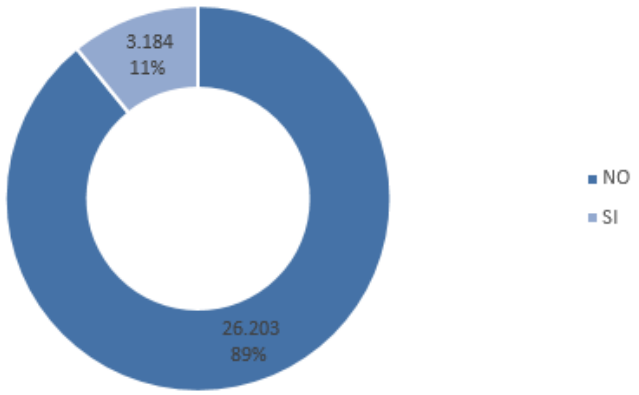

Figura 11. Distribución de los créditos según si tiene un codeudor asociado.

La Figura 12 presenta el ranking de los motivos de mora más frecuentes expresados por los clientes. El top 5 de estos motivos son ingresos insuficientes, no le han realizado pago de nómina, descuido y/o olvido, prioridad en otros gastos y los que no tienen voluntad de pagar.

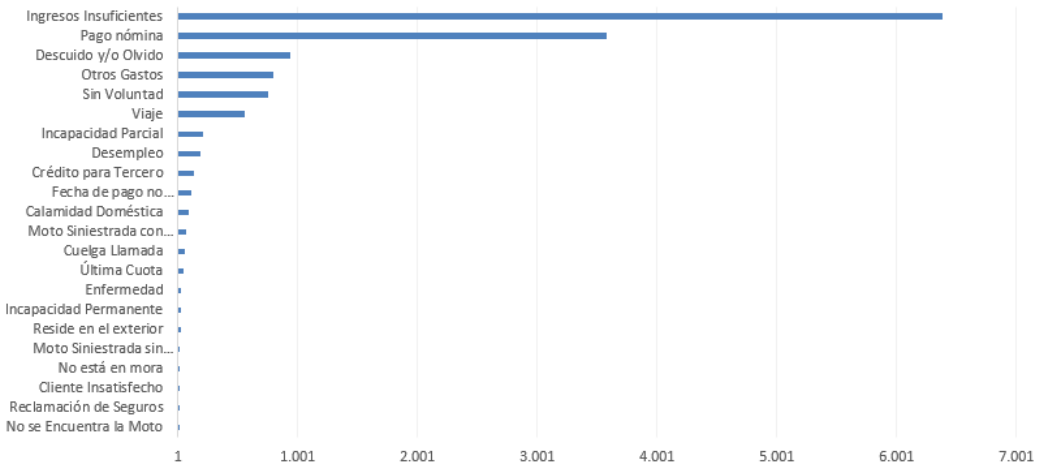

Figura 12. Ranking de los motivos de mora más frecuentes.

A partir de la exploración y comprensión de los datos, se obtiene información valiosa sobre la composición de la cartera de clientes de la empresa. Entender que Bogotá es la zona más grande en cuanto a créditos activos y que a su vez es la zona con menor cantidad de clientes ilocalizados, o que, solo para el 11% de los clientes tenemos una herramienta potente de cobro como lo es codeudor, entender cómo se mueve la cartera por edad de mora, es lo que genera valor en una comprensión de los datos como la presentada.

5. Revisión sistemática de literatura

En el presente trabajo se aplicó la metodología de revisión sistemática de literatura con el objetivo de identificar los aportes teóricos más relevantes en el desarrollo de modelos estadísticos o de aprendizaje automático, para la clasificación de una cartera crediticia, con el fin de mejorar la gestión de cobranza. Técnicas estadísticas y de aprendizaje automático se han propuesto a lo largo de los últimos años, para la calificación crediticia, técnicas que van desde la tradicionales hasta las últimas tendencias y el uso de múltiples clasificadores.

A pesar de que se recopilaron miles de artículos que tratan sobre la calificación crediticia de clientes, la mayoría se basa en la calificación al momento del otorgamiento del crédito y no cuando el sujeto ya es cliente y tiene una obligación que pagar a la entidad. Existe una notable insuficiencia de la aplicación de todas estas técnicas a la gestión de cobranza, para determinar cuáles clientes seguirán siendo morosos y cuales tienen una probabilidad de recuperarse, encontrando una oportunidad de aplicación. Identificada esta problemática, se plantean las siguientes preguntas de investigación:

Q1: ¿Cuáles son las técnicas de aprendizaje estadístico o aprendizaje automático más exitosas aplicadas a la gestión de cobranza?

Q2: ¿Qué porcentaje de los artículos recuperados están enfocados al área de cobranza?

Q3: ¿Qué variables se ha usado en los estudios realizados?

Q4: ¿En qué tipos de negocio se han aplicado estas técnicas?

Q5: ¿Cuáles son los hallazgos recientes basados en el uso de aprendizaje estadístico o aprendizaje automático sobre la calificación crediticia?

Con esta revisión sistemática se busca asentar el trabajo en la mejor evidencia teórica y empírica disponible sobre la gestión de cobro, siguiendo un proceso riguroso para la identificación y evaluación de la misma.

5.1 Metodología

5.1.1 Proceso de búsqueda

Se realizó el proceso de búsqueda a través de la base de datos bibliográficos scopus, en julio de 2019. El tiempo cubierto por la información se delimitó desde 2009 hasta 2019, adicionalmente solo se tomaron artículos de las áreas de economía, econometría y finanza, matemáticas, ciencias de la computación, ingeniería, ciencias sociales, ciencias de la decisión, negocio y gerencia. La cadena de búsqueda utilizada fue:

> (TITLE-ABS-KEY (("debt collec*" OR (collect* W/2 scor*) OR (portfolio W/2 scor*) OR "behavioural scor*" OR ensemble OR "balanced scor*" OR "machine learning" OR "survival analysis") AND ("bad debt" OR "debt collec*" OR portfolio OR "classif* customer" OR "loan risk" OR default* OR prepayment)) AND NOT TITLE-ABS-KEY ("portfolio selection" OR studen* OR medicin* OR biomedic* OR clinic* OR bankruptcy))

5.1.2 Criterios de inclusión y exclusión

Se incluyeron artículos donde se satisfacen los siguientes criterios:

- El artículo presenta metodologías de predicción de incumplimiento de pago que llevan a la segmentación de clientes.
- Los modelos realizados son en el área de cobranza.
- El artículo presenta detalles relevantes del método aplicado.

- El artículo aporta nuevas metodologías o nuevas formas de enfocar el problema.

Se excluyeron artículos donde se cumplen los siguientes criterios:

- El artículo presenta metodologías aplicadas a empresas o sistemas, y no a clientes.
- El artículo presenta metodologías aplicadas para el riesgo de quiebra o pérdidas de la entidad.
- El artículo se basaba en la detección de fraudes financieros.
- El artículo aplica metodologías ampliamente difundidas y no presenta resultados novedosos en su parte práctica o en su parte teórica.

5.1.3 Información recopilada

En este trabajo se recopilaron datos de cada artículo:

- Autores.
- Título.
- Resumen del artículo.
- Nombre de la revista donde se publicó el artículo.
- Objetivos de investigación.
- Resultados obtenidos sobre la metodología en consideración.
- El año de publicación.
- Número de citaciones.

5.2 Resultados de la búsqueda

Con la cadena de búsqueda realizada se recopiló de forma automática 1.549 artículos, que aplicando los criterios de inclusión y exclusión mencionados, finalmente, quedaron seleccionados 23 de ellos cuyas referencias aparecen en la Tabla **2**. Vale la pena resaltar que aunque hay un número alto de publicaciones sobre la calificación crediticia, son muy pocas las que están enfocadas en el área de cobranza.

Tabla 2. Descripción de los artículos seleccionados.

Autores	Titulo	Año	Citas	Metodología/Técnica
Bellotti & Crook,	Credit scoring with macroeconomic variables using survival analysis	2009	72	Análisis de supervivencia usando variables macroeconómicas
Abe et al.,	Optimizing debt collections using constrained reinforcement learning	2010	29	Cadenas de markov para la optimización del cobro de deudas
Sarlija et al.,	Comparison procedure of predicting the time to default in behavioural scoring	2009	19	Análisis de supervivencia con redes neuronales
Bastos,	Ensemble Predictions of Recovery Rates	2014	10	Modelo de ensamble para predecir las tasas de recuperación de las deudas incumplidas
Kim & Kang,	Late payment prediction models for fair allocation of customer contact lists to call center agents	2016	6	Enfoque híbrido para capturar los pagadores tardíos
Dirick, Claeskens, & Baesens,	Time to default in credit scoring using survival analysis: A benchmark study	2017	5	métodos basados en splines y el modelo de cura de mezcla de evento único
Dressler & Stokes,	Survival analysis and mortgage termination at AgChoice ACA	2010	4	Análisis de supervivencia con covariables potencialmente no estacionarias
Kajdanowicz, Plamowski, Kazienko, & Indyk,	Transfer learning approach to Debt portfolio appraisal	2012	2	Transferencia de aprendizaje y aprendizaje basado en métodos de similitud

Tabla 2. (Continuación).

Autores	Titulo	Año	Citas	Metodología/Técnica
Dirick, Bellotti, Claeskens, & Baesens,	Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariates in Mixture Cure Models	2019	1	Modelo de curado de mezcla para incluir covariables que varían con el tiempo
Onar et al.,	A fuzzy rule based inference system for early debt collection	2018	1	Sistema de inferencia difusa para determinar el canal de comunicación más conveniente y el perfil de actividad de comunicación
Wijewardhana,	A mathematical model for predicting debt repayment: A technical note	2018	1	Modelo de regresión logística, redes neuronales artificiales y análisis de afinidad
Jiang, Wang, & Zhao,	A prediction-driven mixture cure model and its application in credit scoring	2019		Análisis de supervivencia, riesgos dependientes del tiempo
Attigeri, M, & Pai,	Framework to predict NPA/Willful defaults in corporate loans: A big data approach	2019		Red neuronal con un componente no estructurado generando datos sintéticos
So, Mues, de Almeida Filho, & Thomas,	Debtor level collection operations using Bayesian dynamic programming	2019		Actualización bayesiana para encontrar una política óptima de qué acción emprender en el próximo período
Li, Li, & Li,	What factors are influencing credit card customer's default behavior in China? A study based on survival analysis	2019		Análisis de supervivencia, modelos de riesgos proporcionales de Cox
Górajski, Serwa, & Wośko,	Measuring expected time to default under stress conditions for corporate loans	2019		Matrices de transición de probabilidad de Markov
Shetty & Manoj,R,	Prediction of default credit card users using data mining techniques	2019		Máquinas de vectores de soporte usando smote y adasyn
Cao, Downes, Khan, Wong, & Xu,	Taxpayer Behavior Prediction in SMS Campaigns	2019		Modelo de aumento de Gradiente
Yurynets, Yurynets, Dosyn, & Kis,	Risk assessment technology of crediting with the use of logistic regression model	2019		regresión logística dirigida a estimar la solvencia del cliente
Boz, Gunnec, Birbil, & Öztürk,	Reassessment and Monitoring of Loan Applications with Machine Learning	2018		Uso de diversos modelos de Aprendizaje de máquinas

Tabla 2. (Continuación).

utores	Titulo	Año	Citas	Metodología/Técnica
Cheng & Cirillo,	A reinforced urn process modeling of recovery rates and recovery times	2018		Análisis de supervivencia, Proceso de Urna Reforzada de Recuperación
Marimo, Clive Malwandla, & Gerbrand Breed,	A cross-sectional survival analysis regression model with applications to consumer credit risk	2017		Enfoque de regresión logística para el análisis de supervivencia
Sun, Wiering, & Petkov,	Classification system for mortgage arrear management	2014		Bosques aleatorios equilibrados

5.3 Discusión

5.3.1 Q1: ¿Cuáles son las técnicas de aprendizaje estadístico o aprendizaje automático más exitosas aplicadas a la gestión de cobranza?

En los últimos 10 años se han aplicado numerosas técnicas para la clasificación de clientes y la gestión de cobro, algunas de ellas como, redes neuronales, regresión logística, modelos difusos, diversos modelos de análisis de supervivencia, máquinas de vectores de soporte, bosques aleatorios, árboles de clasificación, y modelos híbridos, los cuales aparecen relacionados en la Tabla **2**. En general, se puede observar, que en esta área hay más aplicaciones del análisis de supervivencia que de técnicas de aprendizaje automático.

5.3.2 Q2: ¿Qué porcentaje de los artículos recuperados están enfocados al área de cobranza?

Se encontró que solo el 1,5% de los artículos recuperados, están enfocados en el área de cobranza; adicionalmente, a pesar de que estos artículos estudian métodos de clasificación de clientes según incumplimientos de pagos, y de identificación de clientes morosos, solo 4 artículos desarrollan como se hace más eficiente la gestión de cobros al definir la estrategia a seguir, según el modelo planteando.

En el modelo bayesiano propuesto por So, Mues, de Almeida Filho & Thomas (2019) se optimizan las políticas de cobro lo que permiten que los expertos tomen decisiones de las acciones a aplicar en cada período. Los autores Onar et al. (2018) presentan un sistema de inferencia Difusa con el cual obtienen como resultado determinar el canal de comunicación más conveniente según el tipo de cliente, y el tono en la comunicación según el perfil.

Por otra parte, Kim & Kang (2016) proponen un enfoque hibrido que supera en precisión a los árboles de clasificación; y Sun, Wiering & Petkov (2014) platean un modelo de bosques aleatorios equilibrados; en ambas investigaciones los modelos buscan identificar que perfiles de clientes se retrasan en los pagos, y cuales definitivamente si son perfiles morosos a los cuales se les debe realizar una gestión de cobro más fuerte.

5.3.3 Q3: ¿Qué variables se ha usado en los estudios realizados?

Las variables comúnmente usadas en las investigaciones están referidas a variables sociales y variables de comportamiento crediticio. Resaltan con especial frecuencia: ingresos, estado civil, endeudamiento, valor del crédito, ubicación geográfica, edad, plazo del crédito, pago mensual, indicador de refinanciado, tiempo con el banco,

antigüedad laboral, calificaciones de buró, indicador de garantía, cantidad de pagos, historial de atrasos y motivo de mora; todas estas variables se evidencian en las investigaciones de Sarlija et al. (2009), Górajski, Serwa & Wośko (2019), Shetty & Manoj,R (2019), Kim & Kang (2016), Attigeri, M & Pai (2019) y otros.

Adicional a las variables usadas más frecuentes, Bellotti & Crook (2009) introducen variables macroeconómicas como tasa de interés, precio finca raíz, producción nacional, índice del consumidor y el índice de desempleo. Sun et al. (2014) incluyen otras variables sociodemográficas asociadas a entornos familiares y habitacionales, como número de habitantes del hogar o si ha tenido divorcios; por tratarse de créditos hipotecarios.

5.3.4 Q4: ¿En qué tipos de negocio se han aplicado estas técnicas?

Estas técnicas se han usado en dos grandes ramas del mundo de crédito, en la calificación crediticia, la cual se usa en la etapa de la colocación, teniendo esta la mayor aplicabilidad; y en el comportamiento crediticio en la etapa de cobranza.

Dentro de estos horizontes se encuentran aplicaciones más específicas, Shetty & Manoj,R (2019) y Bellotti & Crook (2009) en tarjetas de crédito, Sun et al. (2014) en créditos hipotecarios, Górajski, Serwa & Wośko (2019) en préstamos corporativos, y Onar et al. (2018) en pagos de facturas.

5.3.5 Q5: ¿Cuáles son los hallazgos recientes basados en el uso de aprendizaje estadístico o aprendizaje automático sobre la calificación crediticia?

Se ha demostrado en los estudios que dichas técnicas son capaces de optimizar la cartera de deudas, estimar los clientes que no pagarán y sugerir cuando y que canal

seguir con cada cliente, pues se han venido trabajando de forma innovadora haciendo uso de otras metodologías que ayudan a potencializar estos resultados.

Se observa que implementan uso de covariables y variables macroeconómicas, balanceo de clases, transferencia de aprendizaje, y en muchos casos modelos híbridos entre técnicas de aprendizaje estadístico y automático, como lo es la investigación de Sarlija et al., (2009) quienes aplican el modelo de análisis de supervivencia con los modelos de redes neuronales de acuerdo con la importancia relativa de las diferentes variables para predecir el tiempo de incumplimiento.

Sun et al. (2014) aplica un modelo de bosques aleatorios balanceados para construir un sistema de clasificación. El sistema resultante se desplegó en el departamento de gestión, y esto llevó a grandes ahorros de costos al banco, pues solo enfocaban las estrategias de cobranza en los clientes que el modelo clasificaba como moroso. Onar et al. (2018) con su sistema de inferencia difusa sugiere cual es el canal y el tono de la comunicación más conveniente para cada tipo de cliente.

Sarlija et al. (2009) aplica el modelo de análisis de supervivencia con los modelos de redes neuronales de acuerdo con la importancia relativa de las diferentes variables para predecir el tiempo de incumplimiento.

Bellotti & Crook (2009) muestra que el análisis de supervivencia es competitivo para la predicción de incumplimiento en comparación con el modelo de regresión logística; por la ventaja que tiene de incluir variables macroeconómicas que varían en el tiempo, las cuales no pueden incluirse fácilmente en los modelos de regresión logística.

Modelos de redes neuronales proporcionan una mejor precisión al ser comparados con modelos de machine learning tradicionales en la investigación de Attigeri, M & Pai (2019). Shetty & Manoj,R (2019) compara algoritmos de equilibrio, como los algoritmos SMOTE y ADASYN, utilizándolos como datos de entrada a un modelo de aprendizaje automático como SVM para predecir el incumplimiento de pago de tarjetas de crédito.

6. Justificación de este trabajo

La empresa Créditos Orbe actualmente no realiza una gestión de cobranza de forma eficiente; por el contrario, todos los meses se ve obligada a entregar cartera a casas de cobro porque la capacidad de los asesores para gestionar el 100% de la cartera con la calidad debida es un trabajo muy difícil. El encontrar un modelo que permita realizar una buena segmentación de los clientes según su probabilidad de pago, tiene un impacto positivo en el proceso de implementación de estrategias de cobro, lo cual a su vez generaría mayor rentabilidad para la empresa. El tener modelos de este tipo apoya a la empresa para tomar decisiones más informadas, y además ayuda a reducir las ineficiencias asociadas a la subjetividad, procesos manuales y falta de información. Este trabajo busca apoyar el proceso de gestión de cobro con base a un mejor modelo de segmentación. Esta es la primera motivación para realizar este trabajo.

Por otra parte, pocas veces los modelos estadísticos o de aprendizaje automático son usados para la creación de estrategias de recaudo, mientras que son muy comunes en la predicción de incumplimientos de pago. Por lo que, este trabajo busca apoyar el proceso de gestión de cobro, con la implementación de un prototipo basado en aprendizaje automático, que logre la segmentación de la cartera. La revisión de literatura muestra que existen evidencias de que los modelos de aprendizaje automático son potencialmente útiles en la gestión de recaudo, ya que sus resultados permiten desarrollar estrategias para minimizar esfuerzos de cobranza y facilitar los procesos de los clientes morosos, incrementando la rentabilidad de la cartera. Este trabajo busca contribuir mediante la evaluación de modelos de aprendizaje automático en la gestión de cobro en un escenario real, para los cuales se han reportado pocas experiencias en la literatura. Esta es la segunda motivación de este trabajo.

7. Hipótesis

Se plantea como hipótesis la siguiente: Es posible desarrollar un modelo de score de cartera que use técnicas de aprendizaje de máquinas, con ejecución de forma semanal, y que logre la segmentación de los clientes según su comportamiento de pago, permitiendo el desarrollo de una mejor estrategia de cobro en Créditos Orbe.

8. Objetivos de investigación

8.1 Objetivo General

Desarrollar un prototipo orientado a segmentar la cartera de clientes de Créditos Orbe de manera automática, que permita desarrollar una mejor estrategia de cobro, mediante ejecuciones de frecuencia semanal.

8.2 Objetivos Específicos

1. Realizar un inventario de modelos alternativos que sean más precisos a la luz de las métricas de calidad
2. Establecer parámetros que permitan determinar cuándo se deben autocalibrar los modelos.
3. Implementar un prototipo con el nuevo modelo estimado basado en el sistema autocalibrado.

9. Inventario de modelos disponibles

Se evaluaran los siguientes modelos:

9.1 Vecino más cercano

En este método, se predice un nuevo punto de datos sobre la base del conjunto de datos que ya está disponible. Para esto, la letra K es el número total de vecinos más cercanos. Si un gran número de vecinos K de la nueva clase pertenecen a una clase particular, entonces otros puntos de datos recién pronosticados también se colocarán en esa clase. La distancia entre cada punto de datos en la clase y el nuevo punto de datos predicho se puede calcular utilizando la fórmula de Manhattan, distancia de Hamming y distancia euclidiana. (Cunningham & Delany, 2014)

9.2 Regresión logística

El concepto matemático central que subyace en la regresión logística, es el logit: el logaritmo natural de una razón de probabilidades. En general, la regresión logística es adecuada para describir y probar hipótesis sobre las relaciones entre una variable de resultado categórica y una o más variables predictoras categóricas o continuas (Peng & Ingersoll, 2014).

9.3 Bosques aleatorios

Es una técnica considerada un caso especial de bagging, pues supera lo problemas de correlación entre los árboles de esta técnica. Consiste en construir árboles no correlacionados entre sí, para lograr mayor reducciones en la varianza. Lo hace obligando a cada división a considerar solo un subconjunto de los predictores

disponibles. Por lo tanto, otros predictores tendrán más oportunidades, y por consiguiente, el promedio de árboles resultantes es menos variable (James, Witten, Hastie & Tibshirani, 2013).

9.4 Voto mayoritario

El modelo propuesto se basa en el esquema de votación por mayoría de clasificación de conjunto sobre ciertos tipos de clasificadores base que son de baja complejidad computacional. Cada uno de los tres clasificadores básicos es un experto en una región diferente del predictor. Los tres clasificadores se combinan de tal manera que produce un clasificador de votación por mayoría de conjunto que sea superior a cualquiera de los reglas individuales (Anagnostopoulos & Skourlas, 2014)

9.5 Adaboost

Es un algoritmo de aprendizaje supervisado diseñado para encontrar un clasificador binario que discrimina entre ejemplos positivos y negativos. Aumenta el rendimiento de clasificación de un algoritmo de aprendizaje simple combinando una colección de clasificadores débiles a un clasificador más fuerte. Un clasificador fuerte generalmente produce buenos resultados. Para impulsar un clasificador débil, se aplica una secuencia de aprendizaje. Después de cada ronda de aprendizaje, los ejemplos se vuelven a ponderar para aumentar la importancia de aquellos que fueron clasificados incorrectamente por el clasificador débil anterior. El clasificador fuerte final toma la forma de un perceptrón, una combinación ponderada de clasificadores débiles seguida de un límite (Stachniss, 2005).

9.6 Bagged con árboles de decisión

Es un procedimiento para reducir la varianza de un método de aprendizaje automático. La forma más simple de entenderlo es, el promedio de un conjunto de

observaciones reduce la varianza. Por lo tanto, una forma natural de reducir la varianza y aumentar la precisión de un método de aprendizaje, es tomar repetidas veces muestras del conjunto de datos de entrenamiento, se construye un modelo de predicción por separado en cada conjunto, y se promedian las predicciones resultantes; obteniendo así un único modelo de aprendizaje de baja varianza (James, Witten, Hastie & Tibshirani, 2013).

10. Calibración y validación de modelos

Dado que el objetivo del trabajo es poder hacer una clasificación de la cartera de Créditos Orbe, asentados en las probabilidades de que el cliente pueda o no pagar, con la gran cantidad de información que se tiene disponible, se considera realizar los modelos descritos en el capítulo anterior teniendo como variable respuesta si el cliente ha pagado consecutivamente su cuota en los últimos 3 meses. Las variables explicativas están orientadas en el comportamiento de pago del cliente y variables sociodemográficas.

El entorno de desarrollo tanto para la limpieza de datos como para el modelado ha sido Python ®. Los datos son descargados desde el aplicativo de SAP y de las bases de datos Oracle de la empresa, donde fueron preparados para el proyecto, dejándolos en el formato definido para la carga por medio de los scripts de Python ®. Previo al modelado, se realiza el pre-procesamiento de los datos, como la limpieza y depuración de los datos, a través de las librerías NumPy y Pandas de Python ®. Tal como se explica en la comprensión de los datos, en esta fase, se eliminaron 18 variables que tenían más del 5% de valores nulos y 1.441 registros correspondientes a datos atípicos y valores nulos.

Luego del pre-procesamiento de datos, se encuentra la amplia cantidad de valores que toman algunas variables continuas, y que pueden limitar la precisión del algoritmo. Por lo que el paso siguiente consiste en convertir las variables categóricas en dummys, y luego, para mitigar el efecto de las diferentes escalas, se estandarizan todas las variables a través de la librería StandardScaler de Python ®.

Una vez estandarizadas las variables, se definió utilizar la técnica de validación cruzada, se particiona 10 veces los datos en entrenamiento y prueba, 80% y 20% respectivamente, y para cada partición se aplican algoritmos diferentes de aprendizaje automático; esto con el propósito de realizar una comparación de los

valores reales y los valores predichos por los modelos y poder obtener medidas sobre la precisión de la predicción. Esta técnica junto con la gran cantidad de datos que se tiene, apuntan directamente a encontrar un ajuste óptimo entre el ajuste insuficiente y el ajuste excesivo.

Con los datos de prueba, se realiza la evaluación del modelo comparando métricas de exactitud, precisión, sensibilidad, f1 score y tiempo de corrida (Tabla 3), para finalmente seleccionar el modelo que tiene una mejora significativa en dichas métricas y con menos tiempo de ejecución. Como esto son procesos iterativos, es posible que al realizar ajustes en cada procedimiento, se debe regresar a tareas anteriores.Dado que el score actual de la empresa tiene una exactitud del 71% utilizando una regresión logística, los interesados esperan como mínimo, el mismo ajuste en el modelo a seleccionar.

Siguiendo el diseño planteado, se muestran los resultados obtenidos vía validación cruzada. La Tabla 3 contiene el resumen de las métricas promedio de cada modelo después de realizar las 10 corridas.

Los modelos con las mejores métricas son los ensambles. El modelo bagged decisión tree es el que tiene mejor comportamiento en cuanto a las métricas, una exactitud de 81% seguido de bosques aleatorios, el cual tiene una exactitud de 79%, sin embargo el tiempo de ejecución de bosques aleatorios es de aproximadamente 1 minuto y de bagged con árboles de clasificación de 15 minutos. El modelo con todas las métricas más bajas es vecino más cercano, y también el tiempo de ejecución más alto.

Tabla 3. Resultados de las corridas de los modelos.

Modelo	Exactitud	Precisión	Sensibilidad	f1_score	Tiempo
KNN	0,656830	0,653156	0,744075	0,674214	1176,17

Tabla 3. (Continuación).

Modelo	Exactitud	Precisión	Sensibilidad	f1_score	Tiempo
Adaboost	0,786714	0,774698	0,824090	0,793826	58,99
Regresión Logística	0,690192	0,678841	0,802946	0,712254	11,72
Random Forest**	0,791132	0,789464	0,809033	0,795065	76,13
Bagged Decision Tree	0,813437	0,784709	0,863492	0,821913	911,43
Adaboost	0,786714	0,774698	0,824090	0,793826	58,99
Voting Ensemble	0,761383	0,766824	0,747804	0,753182	104,84

Al alcanzar las métricas presentadas con los ensambles, se decide no seguir probando con modelos más avanzados y seleccionar los bosques aleatorios como solución final.

11. Recalibración del modelo

Se plantea realizar cada 3 meses un seguimiento, el cual consiste en hacer un backtesting, y así comparar las métricas actuales del prototipo, con las que puedan resultar en dicho seguimiento. Es una señal de alerta si la exactitud del modelo es inferior a 75%, que es el valor esperado por la empresa; y requerirá intervención urgente si la exactitud es inferior a 70%. Los análisis de rentabilidad y efectividad de las estrategias implementadas, acompañaran estos resultados, pues si la estrategia empieza a perder efectividad, se debe pensar en revisar el modelo. Adicionalmente, el área de analítica cada seis meses debe revisar si se han creado indicadores o categorías nuevas de control, para evaluar si estas nuevas variables le aportarían significativamente al modelo y poderlas incluir.

12. Evaluación del negocio

A pesar de que el score actual de la empresa basado en el modelo de regresión logística, tiene un ajuste del 71%, que se considera estadísticamente bueno; el modelo dejó de usarse, según instrucciones de expertos e interesados, porque los resultados de las estrategias no estaban siendo los esperados. Por tal motivo, se hace necesaria la evaluación del prototipo por parte de las personas del negocio, las cuales plantearon unas hipótesis y estrategias para pilotear la salida a productivo del prototipo.

Las Figuras 13 y 14, muestran como se distribuye la cartera de Créditos Orbe, según la probabilidad de pago de los clientes basadas en el prototipo vs la probabilidad de pago del modelo actual.

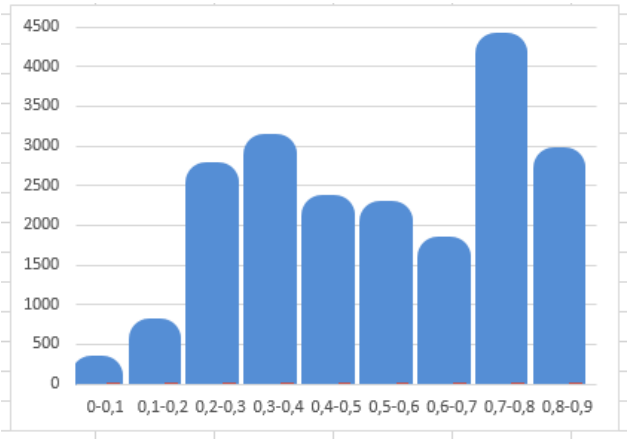

Figura 13. Distribución de la probabilidad de pago de los clientes de Créditos Orbe con los resultados del prototipo.

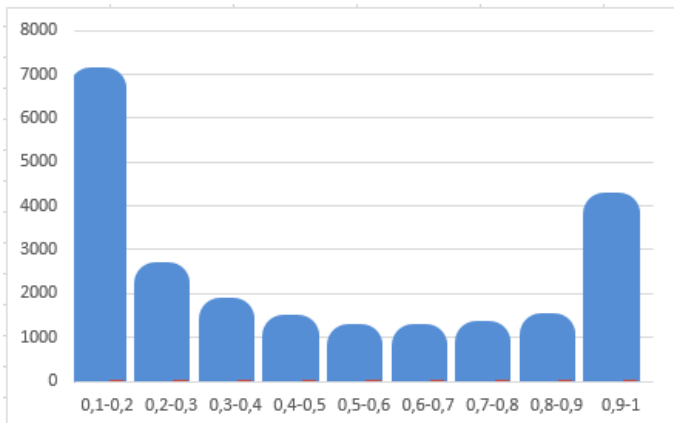

Figura 14. Distribución de la probabilidad de pago de los clientes de Créditos Orbe con los resultados del modelo actual de la empresa.

En validaciones con expertos del área, se discute que la distribución de las probabilidades del prototipo es mejor al compararla con la distribución de las probabilidades del modelo actual. Los interesados y expertos acompañan los hallazgos con el supuesto que al tener tantos clientes que pagan a tiempo sus cuotas, es normal que salga un buen volumen de clientes con altas probabilidades de pago; caso contrario pasa al observar la Figura 14, es muy difícil, según los expertos, que 7.000 clientes de la cartera, esté en un rango de probabilidad de pago de 0,1 a 0,2; adicionalmente que el comportamiento de la distribución en general no les genera confianza.

Para finalizar, la empresa logró desplegar el prototipo el 15 de enero de 2020, acompañados de unas estrategias de cobro orientadas a pilotear el modelo, con el fin de probar que se podía ser más eficiente con la gestión de cobranza si el modelo lograba segmentar bien a los clientes.

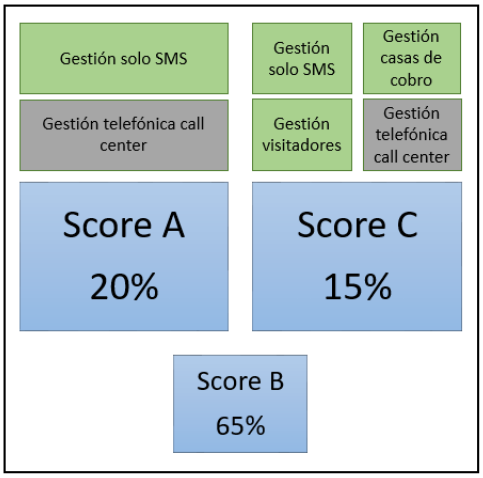

Figura 15. Estrategias de cobro por cada segmento según el prototipo.

Tal segmentación fue: clientes con alta probabilidad de pagar (Score A), clientes con probabilidad media de pago (Score B) y clientes con baja probabilidad de pago (Score C). En la Figura 15 se evidencia las estrategias que se llevaron a cabo con cada segmento, los expertos e interesados construyeron hipótesis y decidieron hacer las primeras pruebas solo con los clientes de los extremos, es decir, Score A y Score C; los clientes Score B se les hizo gestión desde el call center a través de llamadas, convirtiéndose esta estrategia en los grupos de control de los clientes Score A y C.

La hipótesis a probar con los clientes Score A fue, independiente del canal el cliente igual pagará, este segmento se dividió dos grupos iguales de forma aleatoria, un grupo solo se gestionó a través de mensajes de texto y el otro grupo denominado control, fue entregado al call center para gestión telefónica. La hipótesis a probar con los clientes Score C fue, no importa el tono o el canal el cliente no pagará, este segmento se dividió en cuatro grupos iguales de forma aleatoria, el primer grupo solo se gestionó a través de mensajes de texto, el segundo grupo se dividió en partes iguales y se entregó a dos casas de cobro, el tercer grupo se envió al área de

visitadores de la empresa, y el cuarto grupo denominado control se entregó al call center para gestión telefónica (Figura 15).

Al analizar los resultados de las estrategias, mostrados en la Tabla 4, para el Score A, la efectividad de los dos grupos es igual, para el score C, las estrategias con mejor efectividad son las de las casas de cobro con un 20% y 17%, después de le sigue la estrategias de los visitadores con un 15% de efectividad, y por último quedan las estrategias de mensaje de texto y las llamadas desde el call center con un 12%.

Tabla 4. Efectividades de las estrategias propuestas.

	Estrategia	Efectividad
Score A	Mensaje de texto	48%
	Asignación call center	49%
Score C	Mensaje de texto	12%
	Casa de cobro 1	17%
	Casa de cobro 2	20%
	Visitadores	15%
	Asignación call center	12%

En un debate de los resultados con el área de analítica, expertos e interesados, se concluye que los clientes Score A, de igual forma van a pagar su cuota, sea porque se les haga una llamada o se les envíe un mensaje de texto, y es más rentable para la empresa enviarles mensajes de textos a todos estos clientes. Para los clientes Score C, se constata que son clientes más difíciles para cobrar y la estrategia más rentable para la empresa con este tipo de segmento, es entregarlos a casas de cobro

expertas en cobros agresivos. En el mes de febrero los resultados de las estrategias se mantuvieron, por lo que para el mes de marzo la empresa tomó la decisión de dejar de hacer grupos de control y gestionar todos esos clientes según la estrategia más efectiva para cada grupo, dado esto, el call center se encargará de gestionar solo el 65% de la cartera, convirtiéndose esto en ahorros monetarios para la empresa.

13. Distribución

Se le hace entrega al área de analítica de la empresa dos libros de Jupyter ® (Figura 16), el primero con el prototipo para correr en productivo, dicho modelo se correrá una vez por semana y se cargará al sistema de la empresa la segmentación de los clientes. Una vez en el sistema, se establecerán estrategias de tono y canales cobro. Todos los meses se hacen pilotos de estrategias a implementar según la clasificación de los clientes, algunas estrategias que se empezaran a probar son mensajes de texto la primera semana, si en esa semana no han pagado empieza el cobro telefónico; WhatsApp, robots, correos; apuntando así a tener un proceso más eficiente. Con estos pilotos el área de analítica hace análisis de efectividad en el recaudo y rentabilidad de los canales.

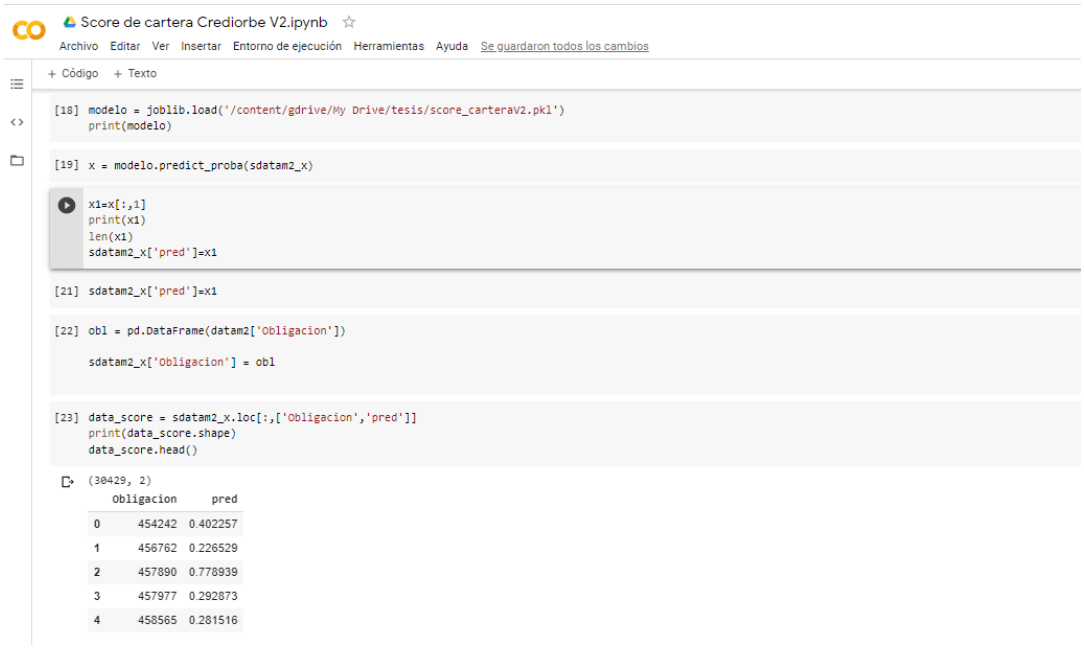

Figura 16. Libro de Jupyter ® para correr el prototipo.

El segundo libro está enfocado en hacer el seguimiento del modelo (Figura 17), el cual se realizará cada 3 meses según lo acordado, haciendo un backtesting, la herramienta muestra las métricas evaluadas de los modelos para entender si el modelo ha mantenido su nivel de exactitud o si por el contrario se ha venido degradando en el tiempo.

Score cartera V2 alertas.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se guardaron todos los cambios

+ Código + Texto

```
tiempo_inicial = time()
abc = AdaBoostClassifier(n_estimators=50, random_state=1)
cv = KFold(n_splits=10, random_state=42, shuffle=False)
#bd_abc = pd.DataFrame({"Modelo":[], "Exactitud":[], "Precision":[], "Sensibilidad":[], "f1_Score":[] })
for train_index, test_index in cv.split(datax):
    print("Train Index: ", train_index, "\n")
    print("Test Index: ", test_index)

    X_train, X_test, y_train, y_test = datax.iloc[train_index], datax.iloc[test_index], datay.iloc[train_index], datay.iloc[test_index]
    abc.fit(X_train, y_train)
    Modelos = Modelos.append({"Modelo":"AdaBoost",
                              "Exactitud":accuracy_score(y_test, abc.predict(X_test)),
                              "Precision":precision_score(y_test, abc.predict(X_test)),
                              "Sensibilidad":recall_score(y_test, abc.predict(X_test)),
                              "f1_Score":f1_score(y_test, abc.predict(X_test))}, ignore_index = True)

tiempo_final = time()

tiempo_ejecucion = tiempo_final - tiempo_inicial

print ('El tiempo de ejecucion fue:',tiempo_ejecucion) #En segundos
```

```
[65] Modelos.groupby(['Modelo']).mean()
```

Modelo	Exactitud	Precision	Sensibilidad	f1_Score
AdaBoost	0.800330	0.777517	0.872509	0.822084
Bagging	0.805257	0.794425	0.853417	0.822593
KNN	0.696661	0.691787	0.773769	0.729909
Logistic Reg	0.740286	0.734003	0.801486	0.765758
Random Forest	0.794310	0.785170	0.842500	0.812583
Voting	0.775873	0.785034	0.796216	0.790033

Figura 17. Libro de Jupyter ® para evaluar las métricas del modelo.

14. Conclusiones

14.1 Objetivos

14.1.1 Objetivo 1: Evaluar si existen modelos alternativos que sean más precisos a la luz de las métricas de calidad

En este trabajo final se introducen distintos tipos de modelos de clasificación existentes, y las metodologías más utilizadas. En los resultados de la revisión de literatura se aprecia que existen metodologías, las cuales superan en métricas de precisión de la regresión logística, el cual es el modelo actual de la empresa. Evaluando los resultados se puede concluir que existen modelos alternativos a la regresión logística que ofrecen mayor precisión y ajuste.

14.1.2 Objetivo 2: Establecer parámetros que permitan determinar cuándo se deben autocalibrar los modelos.

Al área de analítica de la empresa se le hace entrega de un libro de jupyter ® en el cual podrán hacer seguimiento a los parámetros del modelo cada 3 meses. En dicho libro podrán evaluar a través de un backtesting como es el comportamiento del modelo, a través de unos umbrales de rendimiento establecidos, donde les alertará si se mantiene estable, o si por el contrario, ha bajado sus métricas. Los parámetros establecidos fueron los siguientes: exactitud, precisión, sensibilidad, f1 score.

14.1.3 Objetivo 3: Implementar un prototipo con el nuevo modelo estimado basado en el sistema autocalibrado.

En este trabajo final se desarrolló un modelo de clasificación con la metodología de bosques aleatorios; dicho modelo fue seleccionado por tener mejores métricas de exactitud y precisión, además es un modelo eficiente en tiempos de ejecución. Todo el prototipo está dispuesto en un libro de Jupyter ® para que sea ejecutado una vez por semana y lograr la segmentación de la cartera para la implementación de las estrategias que se definan. Este prototipo le permitirá a la empresa ser más eficiente con el uso de sus recursos al momento de la cobranza.

Se puede decir, que aunque muchas entidades de cobro esperan a que sus clientes se encuentren en periodo de mora para tomar acciones sobre ello y estas acciones radican en que los analistas de cobro analicen y llamen a todos los clientes, con el uso de estos prototipos, se ha evidenciado que es mucho más beneficioso conocer de antemano aquellos clientes que son propensos a rodarse de mora para enfocar los recursos en ellos.

14.2 Evaluación de la Hipótesis

El trabajo plantea si es posible desarrollar un modelo de cartera que use técnicas de aprendizaje de máquinas, con ejecución semanal, que logre la segmentación de los clientes según su comportamiento de pago, permitiendo realizar una mejor estrategia de cobro. El prototipo desarrollado fue implementado en enero del presente año, y se ejecuta una vez a la semana, originando una segmentación semanal de la cartera.

Con la implementación del modelo de este trabajo final, y las estrategias implementadas a cada segmento de riesgo identificado por el modelo, se logró reducir la gestión telefónica de los agentes del call en un 65%, pues el segmento de los clientes con alta probabilidad de pagar se gestiona a través de mensajes de texto.

Los clientes con baja probabilidad de pagar son entregados a las casas de cobro, los cuales son expertos en cobros más agresivos. Cada uno de estos canales seleccionados se pilotearon, y resultaron ser los más rentables para la empresa, trayendo esto adicionalmente unos ahorros económicos a la empresa.

15. Referencias

Abe, N., Kowalczyk, M., Domick, M., Gardinier, T., Melville, P., Pendus, C., … Cooley, B. R. (2010). Optimizing debt collections using constrained reinforcement learning. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*, 75. https://doi.org/10.1145/1835804.1835817

Anagnostopoulos, T. T., & Skourlas, C. (2014). *Ensemble Majority Voting Classifier for Speech Emotion Recognition and Prediction*. (August). https://doi.org/10.1108/JSIT-01-2014-0009

Attigeri, G., M, M. P. M., & Pai, R. M. (2019). Framework to predict NPA/Willful defaults in corporate loans: a big data approach. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(5), 3786–3797. https://doi.org/10.11591/IJECE.V9I5.PP%P

Bastos, J. A. (2014). Ensemble Predictions of Recovery Rates. *Journal of Financial Services Research*, *46*(2), 177–193. https://doi.org/10.1007/s10693-013-0165-3

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*(12), 1699–1707. https://doi.org/10.1057/jors.2008.130

Boz, Z., Gunnec, D., Birbil, S. I., & Öztürk, M. K. (2018). Reassessment and Monitoring of Loan Applications with Machine Learning. *Applied Artificial Intelligence*, *32*(9–10), 939–955. https://doi.org/10.1080/08839514.2018.1525517

Cao, G., Downes, A., Khan, S., Wong, W., & Xu, G. (2018). Taxpayer Behavior Prediction in SMS Campaigns. *2018 5th International Conference on Behavioral,*

Economic, and Socio-Cultural Computing (BESC), 19–23. https://doi.org/10.1109/BESC.2018.8697317

Chavarín Rodríguez, R. (2015). Revista mexicana de economía y finanzas. In *Revista mexicana de economía y finanzas* (Vol. 10). Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462015000100073

Cheng, D., & Cirillo, P. (2018). A reinforced urn process modeling of recovery rates and recovery times. *Journal of Banking & Finance*, *96*, 1–17. https://doi.org/10.1016/j.jbankfin.2018.08.014

Cunningham, P., & Delany, S. J. (2014). *k-Nearest neighbour classifiers k -Nearest Neighbour Classifiers*. (April 2007).

Dirick, L., Bellotti, T., Claeskens, G., & Baesens, B. (2019). Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariates in Mixture Cure Models. *Journal of Business & Economic Statistics*, *37*(1), 40–53. https://doi.org/10.1080/07350015.2016.1260471

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, *68*(6), 652–665. https://doi.org/10.1057/s41274-016-0128-9

Dressler, J. B., & Stokes, J. R. (2010). Survival analysis and mortgage termination at AgChoice ACA. *Agricultural Finance Review*, *70*(1), 21–36. https://doi.org/10.1108/00021461011042611

Górajski, M., Serwa, D., & Wośko, Z. (2019). Measuring expected time to default under stress conditions for corporate loans. *Empirical Economics*, *57*(1), 31–52. https://doi.org/10.1007/s00181-018-1435-6

Hector, C. (2011). *Debt Collection in the Information Age : New Technologies and the*

Fair Debt Collection Practices Act Debt Collection in the Information Age : New Technologies and the Fair Debt Collection Practices Act. *99*(6).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning*.

Jiang, C., Wang, Z., & Zhao, H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, *277*(1), 20–31. https://doi.org/10.1016/j.ejor.2019.01.072

Kajdanowicz, T., Plamowski, S., Kazienko, P., & Indyk, W. (2012). *Transfer Learning Approach to Debt Portfolio Appraisal*. https://doi.org/10.1007/978-3-642-28931-6_5

Kim, J., & Kang, P. (2016). Late payment prediction models for fair allocation of customer contact lists to call center agents. *Decision Support Systems*, *85*, 84–101. https://doi.org/10.1016/j.dss.2016.03.002

Li, Y., Li, Y., & Li, Y. (2019). What factors are influencing credit card customer's default behavior in China? A study based on survival analysis. *Physica A: Statistical Mechanics and Its Applications*, *526*, 120861. https://doi.org/10.1016/j.physa.2019.04.097

Marimo, M., Clive Malwandla, M., & Gerbrand Breed, D. (2017). A CROSS-SECTIONAL SURVIVAL ANALYSIS REGRESSION MODEL WITH APPLICATIONS TO CONSUMER CREDIT RISK. In *South African Statist. J* (Vol. 51). Retrieved from https://pdfs.semanticscholar.org/4166/bd323a19ad0256e62e34d64c929610fefa12.pdf

Olivella, V. (2018). *Panorama de las tendencias de crédito en Colombia Segundo trimestre de 2018*.

Onar, S. C., Oztaysi, B., & Kahraman, C. (2018). a Fuzzy Rule Based Inference System for Early Debt Collection. *Technological and Economic Development of Economy*, *24*(5), 1845–1865. https://doi.org/10.3846/20294913.2016.1266409

Peng, C. J., & Ingersoll, G. M. (2014). *An Introduction to Logistic Regression Analysis and Reporting*. (September 2002). https://doi.org/10.1080/00220670209598786

Pontificia Universidad Javeriana. Departamento de Administración de Empresas., M. L., & Saavedra García, M. J. (2010). Cuadernos de administración. In *Cuadernos de Administración* (Vol. 23). Retrieved from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-35922010000100013

Sarlija, N., Bensic, M., & Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications*, *36*(5), 8778–8788. https://doi.org/10.1016/j.eswa.2008.11.042

Shetty, A. S., & Manoj,R. (n.d.). *Prediction Of Default Credit Card Users Using Data Mining Techniques*. Retrieved from https://www.ijitee.org/wp-content/uploads/papers/v8i7/G5754058719.pdf

So, M. C., Mues, C., de Almeida Filho, A. T., & Thomas, L. C. (2019). Debtor level collection operations using Bayesian dynamic programming. *Journal of the Operational Research Society*, *70*(8), 1332–1348. https://doi.org/10.1080/01605682.2018.1506557

Stachniss, C. (2005). *Supervised Learning of Places from Range Data using AdaBoost*. (November 2016). https://doi.org/10.1109/ROBOT.2005.1570363

Sun, Z., Wiering, M. A., & Petkov, N. (2014a). Classification system for mortgage arrear management. *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr)*, (February), 489–496. https://doi.org/10.1109/CIFEr.2014.6924113

Sun, Z., Wiering, M. A., & Petkov, N. (2014b). Classification system for mortgage arrear management. *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 489–496. https://doi.org/10.1109/CIFEr.2014.6924113

Wijewardhana, U. (2018). A Mathematical Model for Predicting Debt Repayment: A Technical Note. *Australasian Accounting, Business and Finance Journal*, *12*(3), 107–115. https://doi.org/10.14453/aabfj.v12i3.8

Wittlinger, B., Carranza, L., & Mori, T. (2008). Mejores Prácticas en Estrategias de Cobranza. *InSight*, *26*(November), 1–23. https://doi.org/10.1177/2158244017731494

Yurynets, R., Yurynets, Z., Dosyn, D., & Kis, Y. (n.d.). *Risk Assessment Technology of Crediting with the Use of Logistic Regression Model*. Retrieved from http://ceur-ws.org/Vol-2362/paper14.pdf