ANALÍTICA DE DATOS APLICADA A LA COBRANZA DE CARTERA

JUAN DAVID MONTOYA YEPES

UNIVERSIDAD EAFIT
ESCUELA DE ADMINISTRACIÓN
MAESTRÍA EN ADMINISTRACIÓN DE NEGOCIOS
MEDELLÍN
2019

ANALÍTICA DE DATOS APLICADA A LA COBRANZA DE CARTERA

JUAN DAVID MONTOYA YEPES[1]

Trabajo de grado para optar por el título de
magíster en Administración de Negocios

Asesora metodológica: Gina María Giraldo Hernández, Ph. D.

Asesora temática: Diana Carolina Parra Giraldo

UNIVERSIDAD EAFIT
ESCUELA DE ADMINISTRACIÓN
MAESTRÍA EN ADMINISTRACIÓN DE NEGOCIOS
MEDELLÍN
2019

[1] dmontoyay@eafit.edu.co

Contenido

Índice de figuras

Índice de tablas

Resumen

Con el fin de mejorar la labor de cobranza de cartera que se realiza diariamente en Cobroactivo S. A. S., la compañía decidió usar los datos ya almacenados para la optimización de sus procesos. Para ello fue necesaria la implementación de modelos de analítica de datos creando un entorno de Big Data[2] que permitiera acceder de forma eficiente a la información mediante una Data Warehouse, además de algunas herramientas para hacer un análisis exploratorio de los datos existentes y evaluar los puntos débiles que se deben mejorar al momento de hacer la gestión del usuario.

Asimismo, se desarrollaron tres modelos de Machine Learning encargados de hacer la segmentación de los deudores, predecir las probabilidades de pago y recomendar de forma óptima cuál es el asesor que se debe asignar a cada deudor y cuál el canal de contacto adecuado para él.

Por último, se desarrollaron dos aplicaciones web. La primera permite el monitoreo de los procesos internos de la compañía automatizando aquellos repetitivos y disminuyendo su tiempo de realización de semanas a segundos; la segunda permite el monitoreo de la labor de cobranza por parte de los clientes y los bancos, ofreciendo un valor agregado.

Palabras claves: Big Data, Machine Learning, análisis exploratorio, modelos predictivos, optimización dinámica, automatización de procesos, recuperación de cartera.

Abstract

In order to improve the portfolio collection work performed daily at Cobroactivo LLC, it was decided to use the data stored in this company to optimize its processes. To do this, it was necessary to implement data analytics models, creating a Big Data environment that would allow efficient access to information through a Data

[2] En este trabajo, los términos (aplicación) Shiny; SQL, (base de datos) SQL, (servidor) SQL; Big Data; Bots, IA Bots; Clustering; Contact Center; Data Quality; Data Warehouse; Dataset; Deploy; Feature Engineering; Forecasting; Machine Learning; Pirme Time; Random Forest; Review; Score (de cobranza preventiva); Bash, (script en) Bash; Spam; Test Data; Trading (algorítmico); Training Data; y Web Scraping, serán usados en el inglés original y sin cursivas.

Warehouse, as well as some tools to make an exploratory analysis of the existing data and evaluate the weak points that must be improved when managing users.

Likewise, three Machine Learning models were developed in charge of debugging the debtors, predicting the payment probabilities and optimally recommending which advisor should be assigned to each debtor and which is the appropriate contact channel for him / her.

Finally, two web applications were developed. The first allows the monitoring of the company's internal processes by automating repetitive processes and decreasing their execution time from weeks to seconds; the second allows the monitoring of collection work by customers and banks, thereby offering added value.

Key words: Big Data, machine learning, exploratory analysis, predictive models, dynamic optimisation, process automation, portfolio recovery.

1. Introducción

En la labor desempeñada por Cobroactivo S. A. S., compañía que presta servicios de gestión integral en recuperación de cartera a los sectores financiero, cooperativo y real y a las cajas de compensación, se ha evidenciado la necesidad de almacenar y procesar la información de manera óptima y eficaz y de garantizar la manipulación correcta del tratamiento de los datos personales de los clientes, aspecto especialmente importante en el negocio de cobranzas que demanda el uso de las mejores tecnologías disponibles en el mercado.

Si se quiere mejorar la calidad del servicio y garantizar la mayor recuperación de cartera posible es necesario ir más allá del simple almacenamiento de los datos de los deudores y pasar a una caracterización activa que permita inferir comportamientos y posibles decisiones y respuestas a diferentes estímulos.

Para ello es necesario entender los factores tanto externos como internos de la compañía, y la única forma de hacerlo es a través de análisis constantes. En el caso de los factores externos, es posible determinar si hay un grupo de usuarios que cumpla ciertas características y esté más predispuesto a pagar: si la ciudad, el barrio en el que se encuentran o el mes en el que se hace el cobro influyen en su voluntad de pago; si hay horarios que sean más pertinentes para el seguimiento de ellos; si la cantidad que deben o las razones de no pago están relacionadas de algún modo con el pago de la deuda; o si se debe perfilar a un deudor y determinar cuál es el descuento mínimo que se le puede ofrecer sobre la deuda para que se anime apagarla.

En el caso de los factores internos, se puede estudiar cuáles son los asesores óptimos para abordar a cierto tipo de clientes; si las horas de contacto influyen; si tienen un mejor rendimiento al hacer la gestión con hombres o con mujeres; si la mejor estrategia para abordar a los clientes es mediante SMS, WhatsApp, llamada o correo; o si para aquellos usuarios de los que no se tiene información actualizada y no ha sido posible contactar se puede encontrar información en las bases de datos del Estado o de otras entidades.

Por todas estas razones se quiere implementar en el entorno de la compañía el uso de nuevas tecnologías como Big Data, inteligencia artificial y Machine Learning (aprendizaje automático).

Para lograr estos objetivos es necesario abandonar el uso de los archivos Excel para guardar la información y pasar a un almacenamiento en la nube en el que se pueda crear un servidor SQL con el que a través de una simple línea de código se pueda extraer información de todas las fechas y clientes que cumplan con ciertas características, sin necesidad de abrir uno a uno cientos de archivos.

Es importante tener en cuenta que no es suficiente con almacenar los datos: a través de técnicas avanzadas como la de Machine Learning, en la cual se enseña a un computador a buscar patrones en datos, se puede llegar a hacer predicciones muy precisas de eventos que aún no se conocen. Para utilizar estos modelos se usan datos de acontecimientos similares pasados para predecir comportamientos futuros.

En el caso específico de Cobroactivo, se decidió que lo mejor era enfocarse en cuatro áreas fundamentales: i) la organización de la información de forma confiable en un entorno de Big Data mediante un Data Warehouse que permite tener acceso inmediato a la información histórica de todas las campañas y sedes; ii) usar esta información para evaluar constantemente cómo va la labor de cobranza y poder tomar acciones correctivas oportunas y a tiempo cuando se está incurriendo en un error; iii) desarrollar modelos de Machine Learning con el fin de realizar la segmentación de los usuarios, calcular la probabilidad de pago y encontrar la forma óptima de asignar asesores y canales de pago a cada deudor; y iv) automatizar los procesos internos que les quitan semanas a la compañía y que, si se automatizan, pueden ser terminados en cuestión de segundos.

2. Metodología

El desarrollo de este proyecto consta de seis fases diferentes, que se desarrollarán de forma secuencial debido a que para su implementación cada una de necesita los resultados de la anterior.

2.1 Fase I. Estudio teórico del problema

1. Recuento histórico de los avances que llevaron al desarrollo de las nuevas metodologías que se aplicarán en el proyecto.
2. Estudio de los conceptos básicos relacionados con Big Data, inteligencia artificial y Machine Learning.
3. Estudio de los conceptos básicos usados en el entorno de recuperación de cartera.
4. Revisión de los métodos usados por otras compañías del mismo sector para la optimización de sus procesos.

2.2 Fase II. Implementación de un entorno de Big Data

En esta fase se van a organizar todos los documentos pertenecientes a la compañía tanto de la sede de Medellín como de la de Bogotá en una única base de datos que permita la extracción eficiente de la información que se va a usar en los futuros modelos.

Esta base de datos se va a desarrollar en un lenguaje que conocido como SQL, que permite acceder a grandes cantidades de información en un mismo lugar, lo que implica que todas las bases de datos pasadas, presentes y futuras de la compañía van a quedar centralizadas de tal forma que mes a mes se pueda hacer la actualización de los modelos y los análisis de optimización correspondientes sin tener que extraer nuevamente los datos de cada uno de los archivos pasados.

Sin duda alguna, esta es la parte fundamental del proyecto, ya que sin ella el análisis de la información se vuelve prácticamente imposible por las siguientes razones:

- La información de cada mes se tiene en archivos Excel que, a medida que aumenta, se vuelven obsoletos.
- Excel no permite centralizar toda la información debido a que solo admite un número máximo de registros.
- Excel no permite hacer análisis estadístico avanzado.
- Excel es un formato que no es compatible con los lenguajes de programación que se van a usar para programar los modelos de optimización y de Machine Learning.
- La actualización de las bases de datos no se puede hacer de forma óptima y se requiere un trabajo manual voluminoso a principios de cada mes.

Para lograr este objetivo es necesaria la implementación de un entorno de Big Data apropiado para la recolección, almacenamiento y análisis de datos relevantes para la compañía a través de los siguientes pasos:

1. Diseño de una base de datos
2. Paso de los datos a un almacenamiento en la nube
3. Organización de datos en la nube
4. Data Quality

2.3 Fase III. Minería de datos

En esta importante fase es donde se va a tener por primera vez una imagen completa de la estructura de los datos en términos de sus dimensiones, estructura y usabilidad.

1. Explorar los datos de forma completa
2. Limpiar los datos
    i. Arreglar los valores faltantes
    ii. Unificar los formatos

3. Analizar posibles sesgos en los datos
4. Comprobar la integridad de los datos

Todo el trabajo de limpieza se hará con el lenguaje de programación R debido a la facilidad en la manipulación de grandes volúmenes de datos y las ventajas que ofrece al haber sido desarrollado como un software estadístico.

2.4 Fase IV. Análisis de los datos existentes

En esta fase se sigue usando el software estadístico R para hacer los primeros análisis de los datos.

1. Hacer un análisis exhaustivo de los datos existentes teniendo en cuenta, entre otros factores, los patrones de pago y las características de los usuarios más propensos a pagar.
2. Determinar cuáles son las áreas en las que puede haber mejoras.
3. Crear tableros interactivos que permitan que la compañía pueda evaluar diariamente su rendimiento y generar alarmas cada vez que algo vaya por mal camino.
4. Automatizar todos los procesos repetitivos de la compañía.

2.5 Fase V. Optimización de los procesos internos

En esta fase, mediante los lenguajes de programación R y Python, se van a construir modelos estocásticos usando la información de las dinámicas internas de la compañía para optimizar las estrategias con las cuales se hace la asignación de los clientes a cada asesor. En otras palabras, se pretende hallar una estrategia óptima que sirva para determinar qué tipo de clientes se le deben asignar a cada asesor haciendo un estudio de la eficiencia pasada de cada uno de ellos.

2.6 Fase VI. Predicción o Forecasting

En esa última fase se van a usar métodos de Machine Learning para entender la estructura interna de los datos y poder predecir cuáles son los clientes que más seguramente van a pagar y que le van a dar más ganancias a la compañía para hacer una labor más fuerte de cobranza y no desperdiciar tanto tiempo con recursos en clientes que muy probablemente no van a proveer ningún beneficio.

Para esta fase se necesita lo siguiente:

1. Hacer un estudio de la estructura interna de los datos.
2. Buscar bases de datos que puedan brindar información adicional de los clientes anteriores.
3. Definir nuevas variables para un estudio óptimo e implementar modelos de clasificación y de Clustering que permitan hacer las predicciones adecuadas.
4. Realizar un proceso de validación cruzada de los modelos para garantizar la predictibilidad.

3. Planteamiento del problema

Cobroactivo S. A. S. es compañía que presta servicios de gestión integral en recuperación de cartera a los sectores financiero, cooperativo y real y a las cajas de compensación, y cuya materia prima para la labor de cobranza son los datos de contacto de los clientes.

Actualmente la compañía se limita a la utilización de los datos de un modo informativo y los almacena en archivos independientes al final de cada mes, en lugar de explotar las capacidades que ellos tienen para influir positivamente en el negocio y mejorar la eficiencia.

Hoy día, pocas empresas de cobranza usan modelos analíticos para mejorar su labor, y las que lo hacer se dedican a la implementación de Bots conversacionales que les permiten reducir la mano de obra. El problema es que dichos Bots son difíciles de entrenar debido al gran número de expresiones coloquiales que tiene el idioma español, por lo que sería más eficiente usar Machine Learning para el mejoramiento de otras áreas, ya que los modelos son más sencillos de entrenar y pueden mejorar positivamente múltiples áreas del negocio.

3.1 Justificación

Debido a la necesidad de mejorar la eficiencia y la productividad de la compañía Cobroactivo S. A. S., es necesario establecer una idea innovadora aplicada al servicio de la cobranza.

Por tal razón se busca desarrollar un modelo basado en la analítica de datos con el fin mejorar la gestión de cobranza, en razón a los beneficios que se pueden lograr cuando los datos se analizan de manera inteligente y ordenada.

3.2 Objetivo general

Proponer un modelo basado en la analítica de datos para un negocio de cobranzas, con el fin de mejorar la productividad de la compañía Cobroactivo S. A. S.

3.3 Objetivos específicos

Analizar las nuevas tendencias a nivel mundial de la analítica de datos aplicada a los servicios de cobranzas.

Desarrollar un entorno de Big Data que permita el almacenamiento de la información en una Data Warehouse de MySQL, con el fin de tener acceso eficiente a los datos.

Desarrollar tres modelos basados en Machine Learning para incrementar y mejorar la gestión y los resultados en un negocio de cobranzas.

Desarrollar una aplicación web que le dé un valor agregado a la compañía permitiéndoles a los clientes monitorear en tiempo real la gestión de la casa de cobranza.

4. Marco conceptual

4.1 Modelos analíticos

La prosperidad y el avance de las organizaciones dependen en gran medida de la forma en que analizan la información de sus clientes. Este panorama se puede agravar si se le suma información algunas veces desconocida: las tendencias del mercado, el panorama competitivo o los cambios en el comportamiento de los consumidores. Los líderes exitosos saben que los conocimientos mejorados pueden marcar la diferencia. El uso de soluciones analíticas dinámicas, construidas para obtener ventajas competitivas, puede discernir las situaciones rápidamente y ajustarlas según sea necesario para tomar decisiones más concretas (TransUnion, 2018).

Las nuevas tecnologías de la información –la internet de las cosas, los datos masivos (Big Data), la computación cognitiva y la realidad virtual– están redefiniendo la sociedad y dinamizando la industria de tal manera que los productos y servicios están cada vez más cerca de las verdaderas necesidades de los clientes y los usuarios. Así lo considera Mark Mattingley-Scott, director de IBM Alemania, para quien el fin de la tecnología es alcanzar un punto de convergencia que permita a las compañías anticipar las necesidades de las personas y proveer los medios para alcanzarlas en el momento y el lugar indicados (Gómez Valencia, 2016).

Es probable que desde el desarrollo de la internet en los años ochenta no exista un término que haya generado tanto revuelo y especulación a nivel mundial como lo ha hecho el Big Data en la última década, más aún si se tiene en cuenta que el 90 % de los datos disponibles en la actualidad fueron creados en los últimos dos años (Marr, 2018).

Hoy día no existe un consenso claro acerca de cómo definir el término *Big Data*, pero generalmente se usa como sinónimo de inteligencia de negocios o de análisis de datos con grandes volúmenes de información (Kitchin, 2014), lo que lleva a la necesidad de diseñar algoritmos y herramientas especiales para su uso (Nadim, 2018). Por esta razón, los sistemas de almacenamiento de información tradicionales como las bases de datos en Access o la manipulación de datos en Excel se quedan cortos (Raia, 2018).

El Big Data nació en el nicho de los negocios *online*, donde se produce un inmenso volumen de datos, a una gran velocidad y de diferentes tipos. Un ejemplo de ello es la compañía estadounidense Amazon, que se dedica a la comercialización de productos. Al ser una de las multinacionales más exitosas a nivel mundial (Fortune 500, 2019), el tráfico que transita por su página web cada día es de millones y millones de usuarios (Statista, 2019), lo que le genera un gran volumen de información en un periodo muy corto de tiempo. Además de que el tipo de datos que se almacenan provienen de diferentes fuentes –desde la identificación del usuario, el número de clics que da en la página, los productos en los que ha mostrado interés, las críticas o reviews de los productos que ha comprado en el pasado, su ubicación, edad o sexo– , en alguna ocasiones, si la búsqueda se hace a través de sus productos Alexa, se almacena también, entre otras, información de voz (Hildenbrand, 2018).

Como fue mencionado, la variedad en la naturaleza de los datos es una de las características principales del Big Data y puede venir de fuentes diferentes.

Datos de la compañía

Son datos específicos que recolecta cada compañía sobre sus clientes, son intrínsecos al negocio y de ellos depende su funcionamiento total. También son su materia prima y generalmente son datos que el cliente proporciona de forma voluntaria al iniciar algún tipo de actividad con ella. Por ejemplo, en el caso de la información que un usuario proporciona a un banco al solicitar un crédito están la identificación, la localización, el estado financiero con el banco y otras entidades, los ingresos, los egresos, etc. (Kitchin, 2014).

Datos web

Son datos que se recolectan a partir del comportamiento de los usuarios en una página web: el número de clics, el tiempo que dura una sesión o los tipos de búsquedas que han hecho en la página, entre otros; uno de los mejores ejemplos es el previamente mencionado Amazon. Estos datos suelen usarse para mejorar el funcionamiento de los motores de búsqueda, hacer ofertas especiales, incentivar las compras, realizar la segmentación de los usuarios, etc. (Kitchin, 2014).

Datos de tiempo y localización

Son datos que recogen a partir de las direcciones IP y GPS, y se producen en todo momento en que haya un usuario conectado a internet, sea desde una tableta, un celular o un computador. Su mayor uso en la actualidad está en las aplicaciones que informan sobre el tráfico: Waze o Google Maps (Kitchin, 2014).

Datos de texto

Son datos que vienen generalmente de cualquier tipo de fuente escrita. Pueden ser individuales si vienen de un correo, una red social, una opinión en un blog o un mensaje de texto; o globales si se toman, por ejemplo, las noticias publicadas por periódicos o revistas. En la actualidad, este tipo de datos son ampliamente utilizados en las campañas políticas, los estudios de mercado y la predicción de conflictos bélicos (Kitchin, 2014).

Datos de sensores

Son datos que provienen de sensores especializados en electrodomésticos, carros, turbinas, etc. Generalmente son los de mejor calidad, ya que su recolección no depende de la intervención humana y son utilizados en la industria para la predicción de fallas en los motores y la optimización de procesos industriales. Comúnmente a estas tecnologías se las conoce como el internet de las cosas, y están revolucionando la forma en que los humanos interaccionan con los artefactos (Oppenheimer, 2014).

Es importante mencionar que actualmente las compañías no se dedican a recolectar un solo tipo de datos, sino que se enfocan en hacer una conexión entre las diferentes fuentes para poder dar una descripción mucho más global de los individuos o de los diferentes mercados (Kitchin, 2014).

Hoy día, un gran número de compañías almacenan sus datos en archivos Excel que luego guardan en carpetas dentro de servidores. Pero, ¿qué tan eficiente es esto? Supóngase que uno de los bancos le proporciona a una compañía un número de cédula y le pregunta en qué mes se hizo la gestión de ese usuario, si pagó o no y de cuánto era la deuda. Teniendo solamente el número de cédula del usuario sería necesario abrir uno a uno todos los archivos Excel de la compañía hasta encontrar el usuario determinado. Este es un trabajo que podría durar días o semanas dependiendo de qué tantos archivos se tenga. De ahí surge la necesidad de encontrar

formas alternativas de almacenamiento que permitan el acceso inmediato a la información.

Esta situación puede resolver creando una base de datos a la que se pueda acceder fácilmente con unas pocas líneas de código. El lenguaje de programación que se usa para este trabajo, llamado SQL (Oracle Corporation, 2019), reduce el problema mencionado a lo siguiente:

```
SELECT cédula, nombre, fecha, deuda, pago
FROM usuarios
WHERE cedula = 12345678
GROUP BY 1, 2, 3, 4, 5
```

Un segmento de código que puede ser ejecutado en solo segundos.

Por más de que sea cierto que los volúmenes de datos son altamente determinantes a la hora de hablar de Big Data, este no es el único elemento de la revolución tecnológica de la actualidad. Su mayor contribución viene con el análisis que se hace de los datos cambiando el paradigma de tomar los ya existentes para entender lo que los usuarios hicieron en el pasado (Moss y Atre, 2003) y usando técnicas como Machine Learning para llevarlo un paso más allá y poder predecir el comportamiento futuro de los usuarios, incluso si son nuevos.

Y es este interés en el futuro y la capacidad de predecir el comportamiento los que hacen del Big Data y el Machine Learning una revolución total, porque permiten a las empresas anticiparse a las fluctuaciones del mercado y evitar comportamientos perjudiciales para los clientes.

Un ejemplo de ello es Netflix, que provee un servicio por suscripción en el que se paga cierto dinero mensualmente para acceder a un amplio catálogo de películas y series. ¿Y cómo usa Netflix el Big Data para mejorar su negocio? Al utilizar Machine Learning desarrolla modelos que les permiten hacer recomendaciones a sus clientes; por ejemplo, la compañía sabe que los usuarios que ven la película A probablemente disfrutarán de la película B, y por eso la recomiendan al inicio de la sesión. Netflix también puede predecir cuándo un usuario se está aburriendo de la plataforma y es probable que quiera cancelar su suscripción, o hacer detección de fraude para evitar que el usuario use tarjetas de crédito fraudulentas en sus transacciones; incluso hace

segmentación de los usuarios según las regiones para la adquisición de contenidos. Este tipo de herramientas, que hasta hace unas décadas eran impensables, son las que han hecho que compañías como Netflix o Spotify tengan tanto éxito a nivel mundial.

Ahora bien, ¿qué es específicamente el Machine Learning y cómo puede ser usado? El Machine Learning (o aprendizaje automático) es un conjunto de técnicas basadas en modelos estadísticos que permiten que los computadores "aprendan" a hacer cosas que para los humanos y animales son naturales: el reconocimiento de patrones y el hecho de poder aprender a partir de la experiencia (Bishop, 2006). Dichos algoritmos de aprendizaje automático emplean métodos de cálculo para "aprender" información directamente de los datos sin depender de una ecuación predeterminada como modelo. Los algoritmos mejoran su rendimiento de forma adaptativa a medida que aumenta el número de muestras disponibles para el aprendizaje. Esto quiere decir que los modelos se actualizan cada vez que hay nuevos datos disponibles (MathWorks, 2019).

Existen dos modelos básicos de aprendizaje con diferentes algoritmos asociados, y se diferencian fundamentalmente en la materia prima que se usa para entrenar el modelo [Figura 1].

El primero, que se conoce como "aprendizaje automático no supervisado", se usa principalmente para descubrir patrones ocultos en los datos. El Clustering, la técnica de aprendizaje no supervisado más común, se emplea para el análisis exploratorio de datos con el fin de encontrar patrones o agrupaciones ocultos en ellos. Entre las aplicaciones del análisis de clústeres están el análisis de secuencias genéticas, la investigación de mercados y el reconocimiento de objetos. Este método usa como materia prima todo el conjunto de datos y no tiene capacidades predictivas, solo exploratorias.

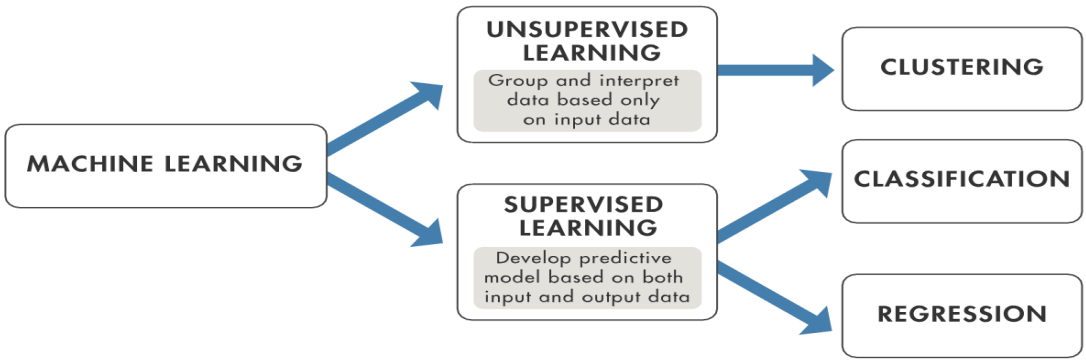
El segundo, que se conoce como "aprendizaje automático supervisado", crea un modelo que realiza predicciones en función de las pruebas en presencia de una incertidumbre. Un algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada y sus respectivas respuestas para estos datos (salidas) y entrena un modelo con el fin de generar predicciones razonables como respuesta a datos nuevos (Bishop, 2006).

El aprendizaje automático supervisado se puede dividir en dos técnicas fundamentales: los algoritmos de clasificación y los algoritmos de regresión.

Las técnicas de clasificación predicen respuestas discretas; por ejemplo, si un correo electrónico es legítimo o es Spam, o bien si un tumor es cancerígeno o benigno. Los modelos de clasificación organizan los datos de entrada en categorías. Las aplicaciones más habituales son las imágenes médicas, el reconocimiento de voz y la calificación crediticia.

Las técnicas de regresión predicen respuestas continuas (Bishop, 2006), por ejemplo, cambios de temperatura o fluctuaciones en la demanda energética, la temperatura en una región o la probabilidad de lluvia. Las aplicaciones más habituales son la predicción de la carga eléctrica y el Trading algorítmico.

Figura 1. Técnicas de aprendizaje automático supervisado y no supervisado



Fuente: MathWorks (2019).

Teniendo en cuenta que ya se conocen los diferentes tipos de modelos que pueden ser usados en los datos, es necesario ahora explorar las herramientas necesarias para la implementación de estos algoritmos.

Así como en el caso del almacenamiento de datos es necesario usar un lenguaje de programación específico que permita acceder a ellos, cuando se va a hacer el análisis y aplicar los modelos de Machine Learning es necesario usar lenguajes de programación que faculten su implementación de forma eficiente.

En teoría, cualquier lenguaje de programación es susceptible de usarse como herramienta para la implementación de modelos de Machine Learning, pero Python y R son los más usados a nivel mundial (Nadim, 2018).

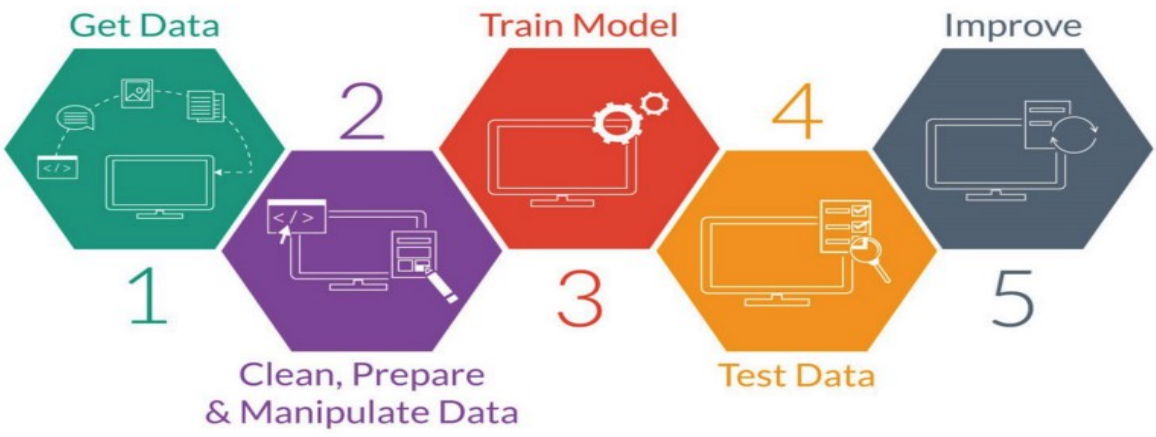
R, un lenguaje de programación dedicado particularmente al análisis estadístico, se desarrolló en 1993 y es uno de los más utilizados en la limpieza y minería de datos debido a las múltiples librerías que tiene desarrolladas para este fin. En la última década, con el auge del Machine Learning, se han creado nuevos paquetes que

permiten hacer este tipo de modelos, aunque su capacidad de cómputo no es muy buena cuando se trata de grandes cantidades de datos (The R Foundation, 2019).

Para superar este problema es necesario que al hacer los modelos de aprendizaje automático se use Python, un lenguaje de programación aparecido en 1991, que es la herramienta perfecta para hacer Big Data y Machine Learning, no solo porque es fácilmente integrable con otros lenguajes como SQL y R, sino por el amplio catálogo de paquetes desarrollados (Pyhton Software Foundation, 2019).

Con la información que se tiene sobre los lenguajes usados y el tipo de algoritmos es posible determinar, dependiendo de su aplicación, cinco pasos primordiales a la hora de implementar estos modelos con datos reales [Figura 2].

Figura 2. Pasos para implementar los modelos de Machine Learning



Fuente: Cook (2018).

El primer paso consiste en obtener los datos que, como se mencionó, pueden venir de fuentes internas o externas de una compañía. El segundo paso consiste en limpiar, preparar y manipular los datos y es muy importante, ya que si los datos no se encuentran en el formato adecuado, el modelo correspondiente que se use sobre ellos no va a funcionar o, simplemente, va a dar resultados erróneos.

Se debe tener en cuenta que si el modelo se va a desarrollar en una sola oportunidad es posible limpiar y preparar los datos cargándolos del archivo original, pero si es algo que se realizará constantemente es necesario cargar los datos en una base de datos SQL, de tal forma que se pueda acceder a ellos en cualquier momento y no sea necesario estar cargando archivo por archivo (Cook, 2018). El proceso de limpiado consiste en hacerse cargo de los valores faltantes, ya sea eliminando la fila

entera o haciendo una imputación, asegurándose de que los valores de cada columna se encuentren en el mismo formato, sea numérico, categórico o de otro tipo.

El proceso de preparación de los datos consiste en determinar si el Dataset está balanceado y determinar la relevancia de las variables en el modelo.

Por último está la manipulación de los datos, donde es necesario decidir si es oportuno agregar nuevas variables y definirlas (Feature Engineering) de tal forma que aporten información adicional al modelo (Cook, 2018).

El tercer paso consiste en definir cuál es el modelo que se va a usar dependiendo de las características de los datos –si son variables categóricas, numéricas, continuas o discretas– y de si lo que se quiere es hacer un análisis exploratorio o un análisis predictivo. Cuando el modelo se define es necesario dividir el Dataset en dos grupos: uno que contenga aproximadamente el 80 % de los datos (Training Data) y otro que contenga el 20 % restante (Test Data). El grupo más grande va a ser usado para entrenar el modelo, es decir, para que el computador aprenda los parámetros internos del Dataset (Bishop, 2006).

En el cuarto paso se procede a comprobar la precisión del modelo al momento de predecir; para esto se toma el modelo entrenado con el Training Data y se le aplica al Test Data para generar un grupo de predicciones que luego se comparan con los valores reales para determinar su precisión (Bishop, 2006).

El quinto paso consiste en tomar ese modelo y, basado en la predicción hecha en el paso anterior, refinarlo para buscar así una mayor precisión. Para hacer esto es posible crear nuevas variables o modificar los parámetros específicos del modelo volviendo a iterar sobre los pasos 3, 4 y 5 hasta que se esté satisfecho con la precisión encontrada. En general, se considera como bastante buenas las precisiones mayores al 80 % (Cook, 2018).

4.2 Aplicaciones en las empresas de cobranza

En lo que al ciclo de crédito se refiere, el foco en la incorporación de IA Bots (o Bots conversacionales), consistentes en programas de computación que utilizan inteligencia artificial (IA) para desarrollar conversaciones con los clientes, hasta ahora se ha concentrado en la digitalización de la experiencia de cliente y en algunos procesos internos referidos a la solicitud de crédito, la captura y la analítica de datos,

y las respuestas sobre requisitos, condiciones, saldos y canales de pago. La etapa de recuperación de cartera vencida continúa lejana de ser sujeta de estos procesos de transformación, con pequeñas actuaciones en las franjas de cobro administrativo y mora temprana, sin un gran sentido de urgencia, con herramientas como, por ejemplo, mensajes de texto y mensajes de voz personalizados para recordatorios de pago y para incrementar la contractilidad (IBR Latam, 2018).

En la actualidad, la aplicación de modelos de Machine Learning e inteligencia artificial se usan de forma incipiente en el negocio de la cobranza, aunque hay un gran potencial de crecimiento en los próximos años. Uno de los puntos más especializados de la inteligencia artificial es el contacto con los usuarios en los Contact Centers orientados a la cobranza de deuda. Los sistemas tienen acceso a las grandes bases de datos y recogen información de cada interacción; además, pueden hacer contacto por diversos canales como páginas web, chat, llamadas de voz o correo electrónico, entre otras (Morales, 2019).

De esta forma, el sistema está en condiciones de obtener información sobre el interlocutor humano –el lugar de residencia, el trabajo, el nivel de ingresos o el historial de crédito–, con lo cual puede ofrecer mejores soluciones para resolver sus deudas, usando las tres herramientas siguientes (Morales, 2019):

- Personalización por medio de las bases de datos: los sistemas de inteligencia artificial tienen acceso inmediato a las grandes bases de datos sobre clientes y usuarios, algo que por su velocidad de procesamiento solo pueden realizar las máquinas.
- Uso de Big Data: al almacenar todas las actividades del usuario en la red se pueden combinar con los datos recogidos en interacciones anteriores.
- Generación de modelos predictivos: las máquinas inteligentes pueden analizar las interacciones con los clientes y deudores para establecer patrones de conducta y generar modelos predictivos que orienten sus propias decisiones.

El arte de un buen administrador de cobranzas está enfocado en la asignación adecuada de los recursos, y es necesario conocer la probabilidad de pago del cliente; con esta información, los administradores de cartera pueden asignar los escasos recursos de manera objetiva, oportuna y asertiva. El modelo predictivo que puede ayudar a determinar la priorización y la administración de dichos recursos es el Score

de cobranza preventiva, que une muchas variables tanto internas como externas, desde las mismas centrales de riesgo hasta la información histórica de gestión y comportamiento del cliente, relacionando causa y efecto y determinando cuándo un cliente bueno pasa a ser un cliente con problemas de pago en los siguientes meses. Toda esta información permite la asignación correcta de los recursos y, por ende, contribuye a generar una buena administración de cobranzas (Martínez, 2014).

Las nuevas herramientas de Big Data y Machine Learning habilitan las empresas de recobro a optimizar sus campañas a través de los Call Centers. Los datos predicen que para la compañía supone un aumento en los ingresos del 30 % y un ahorro de costes de operación del 25 %; además, reducen sustancialmente los litigios por morosidad, el último recurso de las compañías (Bonastre, 2017).

5. Desarrollo del trabajo

5.1 Estructura de los datos

Cobroactivo es una empresa que apenas comienza a incursionar en el manejo óptimo y la utilización de los datos propios del negocio. En la actualidad, cada cliente envía a principios de cada mes un archivo con la cartera que se le asigna a la compañía; estos archivos contienen la información que los bancos poseen de cada uno de los deudores, desde el número de productos que tiene en mora, el número de días de esta, el capital, el tipo de deuda y la información de contacto. De ese momento en adelante el banco envía, tres veces por semana, archivos con la actualización de la deuda de cada uno de esos deudores. Esta acción acumula, en promedio, un número de 24 archivos por campaña por sede, a los cuales se les debe anexar la información de la gestión que los asesores realizan día a día, sea por SMS, correo electrónico, visita a domicilio o llamada telefónica.

En la actualidad, Cobroactivo usa una plataforma que le ayuda a monitorear la gestión de los asesores. Esta plataforma genera a su vez 26 archivos por campaña por sede que se actualizan diariamente, por lo que el almacenamiento y el monitoreo de esta información se vuelve inmanejable a medida que pasa el tiempo.

Con el fin de centralizar dicha información y acceder fácilmente a ella fue necesario desarrollar un entorno de Big Data que permitiera hacer un uso eficiente de la información que se tiene actualmente. Como los datos disponibles son estructurados, es decir, que pueden ser almacenados en tablas, se creó un Data Warehouse en MySQL.

Es importante mencionar que cada tabla es particular para cada campaña, sede y tipo, por lo que fue necesario diseñar alrededor de 140 tablas por campaña, cada una con estructuras y requerimientos especiales. Para esto se creó un Script de R que permite, para cada nueva campaña, introducir archivos Excel de muestra y generar automáticamente la definición del esquema y sus correspondientes tablas.

Después de diseñar y crear el Data Warehouse, era necesario guardar la información en dicha base de datos, cosa que generó varios desafíos. El primero, que la información se encuentra en una plataforma ajena a la compañía y, por lo tanto, cuando estos archivos eran necesitados, la descarga tenía que hacerse de forma

manual, lo que representaba un gasto de tiempo diario significativo para la persona encargada de esta labor.

Para resolver este problema se creó un script en Python usando una técnica conocida como web Scraping, que se encarga de descargar automáticamente la información de las plataformas web. Este script eliminó la necesidad de tener a una persona descargando manualmente todos los archivos de todas las campañas y sedes diariamente y, adicionalmente, redujo el error humano asociado a esta tarea.

Con los datos a disponibles era necesario, entonces, chequear que la información fuera confiable, es decir, que toda ella fuera descargada completamente, con la estructura correcta y que no se perdiera nada en el proceso; para resolver esta situación se desarrolló un código en R que revisa cada uno de los archivos descargados teniendo en cuenta su estructura, el nombre con el que se guarda, la información faltante, etc., y que salta una alarma si hay algún archivo que genere algún problema.

Luego de verificar la integridad de los datos fue necesario hacer un proceso de transformación seguido de un nuevo proceso de verificación. Lo que se buscaba en este paso era transformar los nombres de las columnas para que fueran consistentes, asignar valores a los espacios vacíos, eliminar información redundante y almacenar los datos en un formato adecuado para que la base de datos pudiera aceptarlos. Si este trabajo se hiciera a mano, sería necesario tener a una persona con dedicación exclusiva a esta labor para poder seguir el ritmo diario de la compañía, y los datos antiguos no podrían ser tratados debido a su gran volumen, ya que la persona tendría que ser responsable de la verificación de millones de filas en miles de archivos Excel. Para sobrellevar esto se creó un código en R que procesa toda la información histórica en menos de un minuto y diariamente procesa la nueva información en menos de un segundo.

Con la información procesada se pudo entonces proceder a su carga en la base de datos. Nuevamente, si este proceso se hiciera manualmente, necesitaría una persona que subiera uno a uno los miles de archivos de la compañía, trabajo que implicaría su dedicación completa. Por esta razón se creó un script en Bash que automáticamente sube a la base de datos los archivos después de ser procesados. Este script corre por aproximadamente un (1) segundo todos los días durante la noche subiendo cada uno de los archivos que se descargan diariamente de tal forma que la

base de datos se mantenga actualizada, y genera una alarma si algún archivo no se cargó de forma correcta.

Es importante tener en cuenta que, de todo el proceso, la creación y el mantenimiento de la Data Warehouse son los pasos más importantes, ya que son los que más tiempo requieren y se encargan de proveer la materia prima para cualquier análisis exploratorio o predictivo.

5.2 Inteligencia de negocios

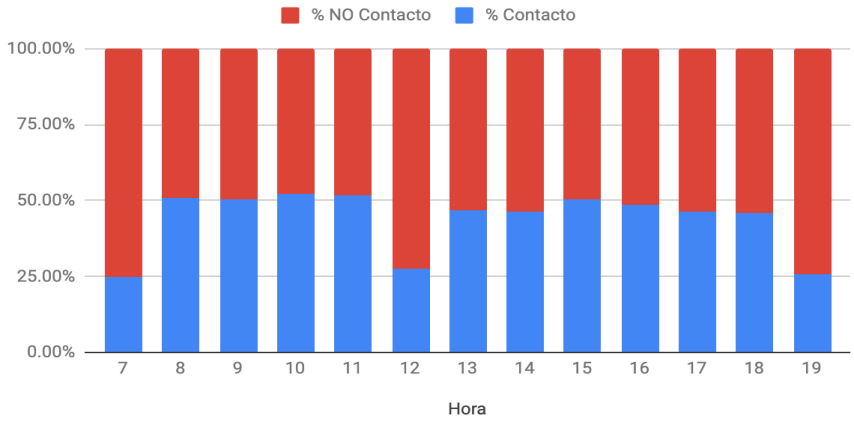
Teniendo todos los datos históricos de forma accesible se procedió a hacer un análisis para determinar si la labor de gestión se ha hecho de forma óptima y cuáles aspectos pueden ser mejorados. Para desarrollarlo se empezó por definir una serie de preguntas que esperaban ser respondidas con los análisis:

- ¿Existe un Prime Time para la gestión telefónica?
- ¿La efectividad del recaudo depende de los días de mora del deudor?
- ¿La efectividad del recaudo depende del tipo de deuda que tiene el deudor (vivienda, libre consumo, etc.)?

Estas preguntas son ejemplos del tipo de cosas que se quieren saber y que pueden usarse para entender mejor el negocio y mejorar la efectividad. Para hacer estos análisis se tomaron datos del Banco de Occidente desde junio de 2016 hasta mayo de 2019 y se encontraron los siguientes resultados:

Se evidenció, por ejemplo, que al momento de hacer gestión telefónica existe un Prime Time, es decir, que hay horarios en los cuales la gestión es significativamente más efectiva. Entre las 7 y las 8 de la mañana, entre las 12 del mediodía y la 1 de la tarde y entre las 7 y las 8 de la noche, la efectividad de contacto disminuye en el 50 % [Figura 3], probablemente porque en esas horas las personas se encuentran fuera de la oficina, ya sea en un medio de transporte o almorzando y, por lo tanto, es más difícil contactarse con ellos, pues muchos de los teléfonos de contacto son números fijos. Esto indica que durante esos horarios los asesores podrían dedicarse a otras tareas como el contacto por correo electrónico o un mensaje de texto, y usar las horas de mayor efectividad para centrarse en la gestión telefónica [Figura 3].

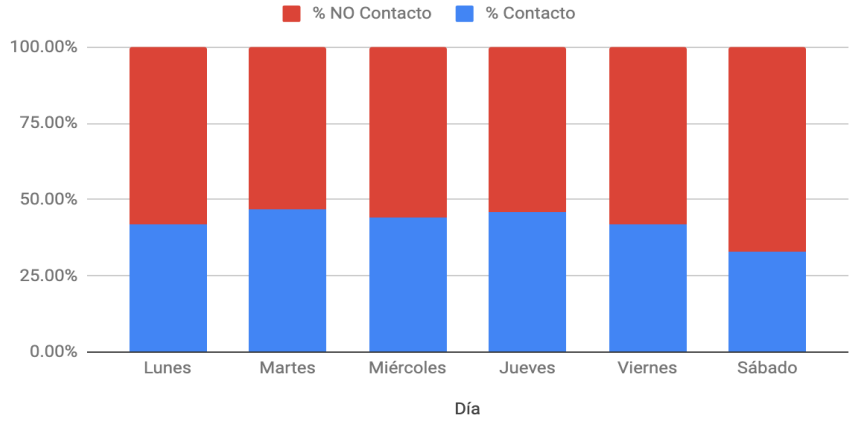
Figura 3. Banco de Occidente. Efectividad del contacto telefónico según la hora del día (campaña junio 2016 – mayo 2019)



Fuente: archivo personal del autor.

Este fenómeno se ve también en los días de la semana y muestra una disminución del 28 % en los sábados comparados con el resto [Figura 4].

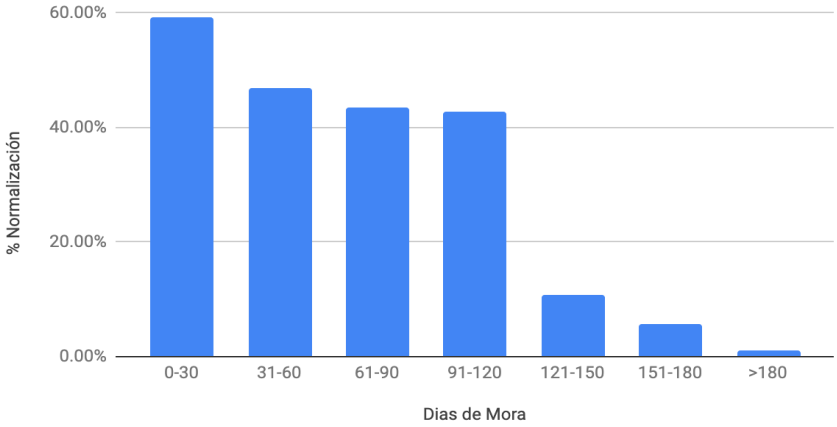
Figura 4. Banco de Occidente. Efectividad del contacto telefónico según el día de la semana (campaña junio 2016 – mayo 2019)



Fuente: archivo personal del autor.

Otra de las características encontradas es que el índice de recaudo disminuye según el número de días de mora, siendo menor del 1 % cuando la mora es mayor a seis meses [Figura 5]. Por esta razón no es conveniente dedicarles mucho tiempo de gestión a estos usuarios, ya que la probabilidad de recaudo es muy baja.

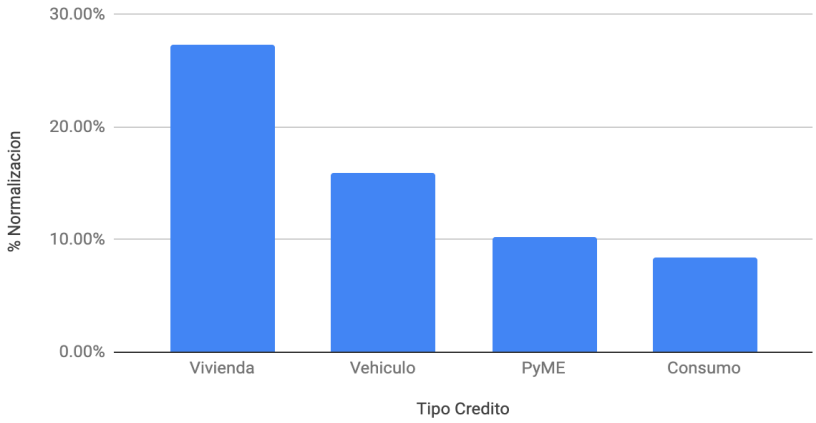
Figura 5. Banco de Occidente. Porcentaje de normalización según el número de días de mora (campaña junio 2016 – mayo 2019)



Fuente: archivo personal del autor.

El porcentaje de normalización para los clientes que tienen créditos de vivienda es de alrededor del 27 %, mientras que el porcentaje para los clientes que tienen créditos de consumo es de alrededor del 9 % [Figura 6.]; esto indica que esta es una variable significativa al momento de determinar la probabilidad de pago.

Figura 6. Banco de Occidente. Porcentaje de normalización según el tipo de crédito (campaña junio 2016 – mayo 2019)



Fuente: archivo personal del autor.

Este tipo de análisis es un primer acercamiento a la inteligencia de negocios aplicada al proceso de cobranza, y a medida que el tiempo avance análisis más profundos y de diferentes rasgos podrán ser realizados. Lo importante es que la infraestructura para hacerlo ya está en pie y que con los resultados de estos primeros datos es posible hacer iteraciones y empezar a responder las preguntas que se vayan generando en el camino.

5.3 Analítica predictiva y Machine Learning

Con toda la materia prima recopilada –los datos– es posible desarrollar modelos de Machine Learning que sirvan como ayuda para mejorar la cobranza. Como primera iteración de esta metodología se hicieron tres modelos diferentes.

El primero utiliza métodos de aprendizaje no supervisado y sirve para hacer la segmentación de los usuarios y determinar las características específicas de cada segmento con el fin de usarlas como herramienta para determinar mejores estrategias de cobranza. Este es un modelo de Clustering que usa el método de K-means, en el que se encuentra que el número óptimo de clústeres es cuatro. Comparado con otros métodos de Clustering, se eligió K-means, porque es el que brinda una mayor interpretabilidad de los segmentos.

El segundo determina la probabilidad de pago de un usuario. En este caso se hizo una comparación entre dos métodos para determinación de probabilidades: una regresión logística y una Random Forest, y se encontró un error de clasificación menor en el caso de la primera, con una precisión cercana al 86,9 %.

El tercero es un sistema de recomendación –un esquema similar al de Netflix o Spotify– que asigna a cada deudor el asesor óptimo y el mejor método de gestión, sea este un SMS, un correo electrónico, un mensaje por WhatsApp, una llamada o una visita.

Para estos tres modelos se siguieron las cinco etapas descritas en el Marco conceptual; la primera y la segunda fueron realizadas al momento de crear la base de datos, y para las etapas 3, 4 y 5 se usaron métodos de validación cruzada y de optimización de hiper-parámetros para asegurar la mejor precisión en los algoritmos. En este momento ya se hizo el Deploy de los modelos y es necesario esperar unos meses para poder evaluar qué tanto ha mejorado la cobranza desde la implementación de estos métodos.

5.4 Automatización de procesos

Uno de los asuntos que presentaban un mayor problema en Cobroactivo es la cantidad de procesos que tenían que hacerse repetitivamente, en razón a que tomaban mucho tiempo, hacían que todo el proceso de cobranza se retrasase y que el monitoreo solo

podía realizarse a final de mes cuando ya no había nada que hacer. ¿Su causa? Los archivos que guardaban la información se mantenían en formato Excel, así que si se quería hacer un análisis era necesario ejecutar cruces manuales entre múltiples archivos, cosa que no solo generaba un gasto de tiempo significativo, sino que también llevaba a muchos errores de tipo humano.

Para mejorar esto se desarrolló una aplicación Shiny, que se puede acceder con un nombre de usuario y una contraseña, y permite hacer un monitoreo diario de las gestiones y los resultados de cada asesor; así, si hay algún segmento que no se está gestionando lo suficientemente o si el asesor se está centrando en los deudores que tienen poca probabilidad de pago, este hecho puede ser corregido inmediatamente.

Esta aplicación hace que los reportes –que en un principio requerían un trabajo de varias semanas de los coordinadores y de los monitores– se creen automáticamente todos los días dependiendo de la gestión diaria y, por consiguiente, el tiempo de los empleados pueda ser invertido en otras cosas que generen valor para la compañía y que, además, su efectividad general mejore, ya que si se está incurriendo en algún error se pueda corregir inmediatamente y no sea necesario esperar hasta que ya no haya nada que hacer.

Por último, con el fin de crear un factor diferenciador entre Cobroactivo y otras casas de cobranza, se pasó todo el contenido a la nube y se creó una aplicación web específica para los clientes que les permite ver cómo va el proceso de cobranza y descargar los reportes de interés. Con esto, los bancos pueden tener una visión más transparente de la forma en la que Cobroactivo hace su gestión y cómo esta evoluciona a lo largo del mes.

5.5 Productos desarrollados

Tabla 1. Cobroactivos S. A. S. Resumen de entregables

Entregables Cobroactivo						
	Tipo	Nombre	Descripción	Tiempo de cómputo	Frecuencia de uso	Uso
1	Data Warehouse	cobroactivo_campanas	Data warehouse en mysql en la cual se tiene un esquema específico para cada campaña.	Solo se crea una vez	Continuo	Automático
2	Código	schema_creation.R	Programa en R que toma archivos de muestra de una campaña y crea su respectivo esquema con tablas personalizadas en la base de datos Cobroactivo_campanas	< 1 s	Cada vez que se necesite añadir una campaña	Semi-automático
3	Código	automatic_file_downloading.py	Programa en Python que descarga automáticamente la información de la plataforma de internet externa	< 1 s	Diaria	Automático
4	Código	data_integrity.R	Script en R que toma los datos descargados por automatic_file_downloading.py y comprueba que no haya información faltante y que no se haya modificado en el proceso de descarga.	< 1 s	Diaria	Automático
5	Código	data_transformation.R	Script en R que toma los datos a los que se les verificó la integridad, los transforma para que queden en un formato adecuado y consistente y guarda una copia de estos para luego ser subidos a la base de datos.	< 1 s	Diaria	Automático
6	Código	automatica_file_loading.sh	Script en Bash que toma los datos transformados y los carga en la base de datos.	~ 1 s	Diaria	Automático
7	Código	analysis.R	Script en R que hace un análisis de los indicadores más relevantes y permite entender cuáles son los patrones de pago de los usuarios y cómo evolucionan a lo largo del tiempo, además de qué es lo que ha pasado en la compañía	~10s	Mensual	Automático

			en años pasados			
8	Código	segmentation.py	Programa en Python que permite hacer una segmentación de los deudores en grupos con características específicas, que permiten entender mejor su comportamiento y características que pueden ser usadas para el mejoramiento de la labor de cobranza.	~ 3 min	Mensual	Automático
9	Código	probabilities.py	Programa en Python que predice la probabilidad de pago de cada deudor y que se actualiza diariamente según los resultados de cada gestión. Este programa presenta una precisión del 85 % y considera una mayor sanción para los deudores que se predice que NO van a pagar y SÍ pagan que para los deudores que se predice que SÍ van a pagar y NO pagan.	~1 min	Diario	Automático
10	Código	assignation.py	Programa en Python que predice la asignación óptima de deudores a cada asesor.	~ 1 min	Mensual	Automático
11	Aplicación web	Informes internos	Aplicación web desarrollada en Shiny que permite que los coordinadores se conecten y vean diariamente cómo va el desarrollo de la gestión, que segmentos no se han gestionado propiamente, si se está invirtiendo el tiempo en procesos que no dan resultados, y si los deudores que tienen mayor probabilidad de pago se han gestionado o no.	Siempre visible	Continuo	Automático

12	Aplicación web	Informes externos	Aplicación web desarrollada en Shiny que permite que los bancos y otro tipo de clientes monitoreen diariamente la gestión y se den cuenta como avanza la gestión, para que tengan un entendimiento más claro de cómo Cobroactivo gestiona los clientes, los canales que usa, la efectividad en la cobranza, etc. Además, permite la descarga automática de informes mensuales que anteriormente tenían que ser creados a mano y ser enviados por correo.	Siempre visible	Continuo	Automático

Fuente: elaboración del autor.

Con todos los resultados obtenidos fue posible la creación de un departamento de analítica de datos en una compañía en la cual estos no estaban siendo utilizados para el mejoramiento de su labor. Se organizó la información en una Data Warehouse alojada en la nube, se dieron los primeros pasos hacia el proceso de hacer inteligencia de negocios, se crearon tres modelos predictivos que usan Machine Learning para hacer más eficiente el proceso de cobranza, se desarrolló una aplicación web que permite el monitoreo de las labores internas de la compañía y la descarga automática de reportes y, por último, se desarrolló de una aplicación web que genera un valor agregado para los clientes, en la cual pueden mirar a lo largo de todo el mes en qué va la labor y cómo la compañía dispone los recursos con que cuenta.

5.6 Progreso

Tabla 2. Cobroactivos S. A. S. Resumen del avance del proyecto

Fase	Paso	Progreso
I	Recuento histórico de los avances que llevaron al desarrollo de las nuevas metodologías que se aplicaran en el proyecto.	Finalizado
	Estudio de los conceptos básicos relacionados con Big Data, inteligencia artificial y Machine Learning.	Finalizado
	Estudio de los conceptos básicos usados en el entorno de recuperación de cartera.	Finalizado
	Revisión de los métodos usados por otras compañías del mismo sector para la optimización de sus procesos.	Finalizado
II	Diseño de una base de datos.	Finalizado
	Paso a un almacenamiento en la nube o servidor.	Finalizado
	Organización de datos en la nube o servidor.	Finalizado
	Data Quality.	Finalizado
III	Explorar los datos de forma completa.	Finalizado
	Limpiar los datos.	Finalizado
	Analizar posibles sesgos en los datos.	Finalizado
	Comprobar la integridad de los datos.	Finalizado
IV	Análisis exhaustivo de los datos existentes hasta ahora teniendo en cuenta los patrones de pago y las características de los usuarios más propensos a pagar, entre otros.	Finalizado
	Determinar cuáles son las áreas en las que puede haber una mejora.	Finalizado
	Crear tableros interactivos que permitan que la compañía pueda evaluar día a día su rendimiento y generar alarmas cada vez que algo vaya por mal camino.	Finalizado
	Automatización de los procesos repetitivos de la compañía.	En curso
V	Hallar una estrategia óptima que sirva para determinar qué tipo de clientes se le deben asignar a cada asesor, haciendo un estudio de la eficiencia pasada de cada uno de ellos.	Finalizado
VI	Hacer un estudio de la estructura interna de los datos.	Finalizado
	Definir nuevas variables para un estudio óptimo de la implementación de modelos de Machine Learning.	Finalizado
	Proceso de validación cruzada de los modelos para garantizar la predictibilidad.	Finalizado

Fuente: elaboración del autor.

6. Conclusiones

Con el fin de poder almacenar de forma adecuada los datos de la compañía, se implementó una Data Warehouse que consta de cuatro esquemas principales, con información de las campañas de Banco de Occidente y AV Villas tanto para la sede Medellín como para la sede Bogotá, y que redujo de horas a segundos el acceso a la información de la compañía.

Se desarrollaron tres modelos de Machine Learning que permiten determinar la probabilidad de pago de un deudor con un 85 % de precisión, el segmento al que el deudor pertenece, los patrones asociados y, por último, la asignación óptima de los deudores a cada asesor. Estos modelos se encuentran actualmente en evaluación y se espera tener resultados en los próximos meses de qué tan útiles han sido para la compañía.

Se creó una aplicación web que le genera un valor agregado a la compañía, ya que le muestra a los clientes, es decir, los bancos, qué labor se está haciendo diariamente y cómo se ha avanzado en el proceso, además de la posibilidad de descargar automáticamente los respectivos reportes. Antes de esto los clientes tenían que esperar al final de cada mes a que se enviara un reporte, y el trabajo realizado día a día por la compañía no era tan transparente.

Por último, se creó una aplicación web que genera automáticamente reportes de los indicadores claves que permiten evaluar diariamente la labor de los asesores de tal forma que los errores que se cometen puedan ser solucionados inmediatamente y no sea necesario esperar a final de mes cuando ya nada se puede hacer.

Después de un estudio exhaustivo de la bibliografía relacionada con la analítica de datos aplicada a las empresas de cobranza y los diferentes productos desarrollados para para mejorar la eficiencia en el proceso de recuperación de cartera, es posible decir que los primeros pasos hacia la aplicación efectiva de modelos analíticos han sido exitosos y lo que se espera en los próximos meses es la evaluación del funcionamiento de los modelos con nuevos deudores, su consiguiente refinamiento y

la profundización en análisis de inteligencia de negocios para complementar lo hecho hasta ahora.

7. Referencias

Banco de Occidente (s. f.). Archivo personal del autor.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Singapur: Springer. Disponible en http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf

Bonastre, R. (2017). *La inteligencia artificial en la gestión de recuperación de deuda* [en línea, 6 de noviembre]. Innovan.do. Disponible en https://innovan.do/2017/11/06/la-inteligencia-artificial-en-la-gestion-de-la-recuperacion-de-deuda/

Chen, C. L. P. y Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314-347, agosto, doi.org/10.1016/j.ins.2014.01.015

Cook, K. (2018). Understand the machine learning from scratch for beginners [en línea, 24 de septiembre]. House of Bots. Disponible en es.mathworks.com/discovery/machine-learning.html

Fortune 500 (2019). *Fortune 500* [en línea]. Disponible en http://fortune.com/fortune500/list/

Gómez Valencia, A. (2016). Big data e internet cambian radicalmente a la sociedad [en línea, 19 de septiembre]. *Agencia de Noticias Universidad EAFIT*. Disponible en http://www.eafit.edu.co/sitionoticias/2016/big-data-internet-cambian-radicalmente-sociedad

Hildenbrand, J. (2018). *Amazon Alexa: What kind of data does Amazon get from me?* [en línea, 27 de marzo]. Androidcentral. Disponible en www.androidcentral.com/amazon-alexa-what-kind-data-does-amazon-get-me

IBR Latam (2018). *Soluciones Multicanal*. Sitio web ibrlatam.com/sitio2018

Ireton, R. (2009). *Computational systems biology*, J. McDermott, R. Samudrala, R. Bumgarner y K. Montgomery (eds.). Nueva York: Humana Press.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Thousand Oaks, CA: Sage.

Lesk, A. (2014). *Introduction to Bioinformatics*. Oxford, Reino Unido: Oxford University Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. y Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Nueva York: McKinsey Global Institute. Disponible en https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

Marr, B. (2018). How much data do we create every day? The mind-blowing stats everyone should read [en línea, 21 de mayo]. *Forbes*. Disponible en https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Martínez, E. (2014) Scoring de cobranza preventiva en la gestión de riesgo [en línea, 5 de febrero]. *Consumer Risk Analytics*. Disponible en https://edgarmartinezq.wordpress.com

MathWorks (2019). *Machine Learning. Tres cosas que es necesario saber* [en línea]. Disponible en es.mathworks.com/discovery/machine-learning.html

Mayer-Schönberger, V. y Cukier, K. (2014). *Big data: A revolution that will transform how we live, work, and think*. Londres: John Murray.

Morales, P. (2019). *Inteligencia artificial en procesos de cobranza de grandes empresas: más allá del Bot* [e-Book]. Providencia, Chile: Digevo Corp. Disponible por descarga en http://soluciones.digevo.com/ebook-inteligencia-artificial-en-procesos-de-cobranza-2

Moss, L. T. y Atre, S. (2003). *Business intelligence roadmap: The complete project lifecycle for decision-support applications*. Boston: Addison-Wesley.

Nadim, J. (2018) Top 5 best programming languages for artificial intelligence field [en línea]. *Geeks for Geeks*. Disponible en https://www.geeksforgeeks.org/top-5-best-programming-languages-for-artificial-intelligence-field/

Oppenheimer, A. (2014) *Crear o morir: cómo reinventarnos y progresar en la era de la innovación*. Barcelona: Debate.

Oracle Corporation (2019). *MySQL 8.0 Reference Manual* [en línea]. Disponible en https://dev.mysql.com/doc/refman/8.0/en/introduction.html

Python Software Foundation (2019). *The Python Language Reference* [en línea]. Disponible en https://www.python.org/

Raia, M. (2018*). 5 things you should stop doing with Microsoft Excel* [en línea, 13 de junio]. Integrify. Disponible en https://www.integrify.com/blog/posts/5-things-you-should-stop-doing-with-microsoft-excel/

Robson, K. (1992). Accounting numbers as “inscription”: Action at a distance and the development of accounting. *Accounting, Organizations and Society*, *17*(7), 685-708, octubre, doi.org/10.1016/0361-3682(92)90019-O

Rouse, M. (2018). *Relational Database* [en línea]. Tech Target. Disponible en https://searchdatamanagement.techtarget.com/definition/relational-database

Statista (2019). Combined desktop and mobile visits to Amazon.com from February to April 2019 (in millions). [en línea]. *Statista*. Disponible en https://www.statista.com/statistics/623566/web-visits-to-amazoncom/

The R Foundation (2019). The R project for statistical computing [en línea]. *The R Foundation*. Disponible en https://www.r-project.org/

TransUnion (2018). Agilice su proceso de toma de decisiones a través del poder de las soluciones analíticas [en línea]. *TransUnion*. Disponible en https://www.transunion.mx/solucion/analiticas

Wang, F.-Y., Zeng, D., Carley, K. M. y Mao, W. (2007). Social computing: From social informatics to social intelligence. *IEEE Intelligent systems*, *22*(2), 79-83, marzo-abril, https://pdfs.semanticscholar.org/f430/9d8913cc9f0d72ec08a4bfb9829866d321d1.pdf

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X. y Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1624-1639, https://ieeexplore.ieee.org/document/5959985