Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Herramienta para la Gestión Empresarial Eficiente: Modelo predictivo de adquisición de préstamos utilizando técnicas de aprendizaje automático

Aplicación en R

AUTORA: DANIELA LAURA ROCCA

SEPTIEMBRE 2019

Resumen

En un contexto de enorme disponibilidad de datos en las organizaciones, mejoras continuas en las capacidades de procesamiento y almacenamiento, herramientas informáticas cada vez más poderosas y un mercado extremadamente competitivo donde los accionistas reclaman mayor certidumbre sobre el rendimiento de sus inversiones, se propone dar respuesta a una problemática organizacional que afecta a la gestión empresarial eficiente y los resultados de una entidad bancaria.

Se desarrollará un modelo predictivo de adquisición de préstamos dirigido al segmento "Nunca Prestamos", el menos rentable. Para ello se aplicarán dos algoritmos de aprendizaje automático de clasificación binaria: *Decision Tree* (*DT*) y *Random Forest* (*RF*). Utilizando como métricas de performance técnica la exactitud y el área bajo la curva ROC.

Bajo un enfoque de ERM[1], donde se busca minimizar los efectos del riesgo en el capital y las ganancias de una organización, se pretende ofrecer una herramienta útil para mitigar el riesgo empresario asumido al llevar adelante este negocio: el riesgo operacional de error en la presupuestación y planificación comercial, dando mayor certidumbre a los accionistas sobre el rendimiento de las inversiones en campañas dirigidas a este universo. El riesgo estratégico de ofrecer el producto adecuado al cliente adecuado[2] y de disponibilidad de capital, liberando capital para su potencial inversión en otras acciones que generan mayor valor para la compañía (costo de oportunidad).

Para ello se pondrá foco en evaluar la conveniencia de la implementación del modelo en base a su impacto en Resultados.

Se demuestra que *RF* posee una performance notoriamente superadora a la de *DT* con una exactitud de 0,956 y área bajo la curva ROC de 0,977. Se evidencia que a partir de la aplicación del modelo se puede obtener una mejora en la ganancia neta anual de 72.4%, y mayor al 300% al considerar el costo de oportunidad del capital disponible.

Se espera sentar las bases para futuros análisis de impacto dinámico y modelos de otros productos.

Palabras clave: Decisiones basadas en datos. Marketing de Big Data. KDD. Minería de datos. Algoritmos de aprendizaje automático de clasificación binaria. Impacto en resultados. ERM.

[1] Gestión de Riesgos Empresariales

[2] Llamado "customer wants" dentro de los posibles riesgos estratégicos definidos por Casualty Actuarial Society en el "Enterprise Risk Management Committee" de mayo de 2003.

Índice

Introducción

Dada la explosión de disponibilidad de las herramientas informáticas de Hardware y Software para el apoyo de los procesos empresariales, hoy día las empresas generan grandes cantidades de información. Hecho que, sumado al aumento de las capacidades de almacenamiento de datos, hacen que las organizaciones puedan disponer de una gran cantidad y variedad de datos relativos a su actividad diaria. Surge allí una gran oportunidad para la resolución de problemas basada en datos reales y ya no en intuiciones o pálpitos personales del grupo dirigente de la empresa. Así, el análisis de los datos toma cada vez una función vital en el proceso de toma de decisiones (Drucker, 1999, pp. 176-177).

Según la visión de la teoría organizacional, se puede definir el conocimiento como la información que posee valor para ella (Stewart, 1999), es decir, aquella información que permite generar acciones asociadas a satisfacer las demandas del mercado (Porter & Millar, 1986) y apoyar las nuevas oportunidades a través de las competencias centrales de la organización. De allí que las organizaciones deben tomar control de sus datos para analizarlos.

Surge entonces la minería de datos como una tecnología y estrategia de modelado matemático que intenta ayudarnos a comprender el contenido de una base de datos. La Minería de datos (*datamining*) es definida por Molina (2019, Pág. 23) como el "proceso de extraer conocimiento de bases de datos" … "descubrir situaciones anómalas y/o interesantes, tendencias, patrones y secuencias en los datos" … "siendo su objetivo construir un modelo a partir de los datos pre-procesados, el cual pueda producir nuevo conocimiento que sea útil para el usuario".

Desde un punto de vista académico, el término minería de datos es una etapa de un proceso mayor llamado *Knowledge Discovery in Databases* (KDD). Se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información (Han & Kamber, 2001). No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Consta de una serie de cinco fases secuenciales que se recorrerán detalladamente en los capítulos 4, 5 y 6.

El número de posibles relaciones es demasiado grande, y resulta prácticamente imposible validar cada una de ellas. Para resolver este problema se utilizan estrategias de búsqueda, extraídas del área de aprendizaje automático (Berry y Linoff, 1997). En función de lo que se busque, deberá emplearse uno de los siguientes 4 enfoques de aprendizaje automático: clasificación, predicción numérica, detección de patrones o agrupación.

La tarea de aprendizaje automático supervisado que se usa con frecuencia para predecir a qué categoría pertenece se conoce como clasificación. Dentro de los modelos de clasificación, algunos algoritmos de aprendizaje supervisado son: Vecino más cercano (*Nearest Neighbor*), Bayes ingenuo (*naive Bayes*), Árboles de decisión (*Decision Trees*), Estudiantes de reglas de clasificación (*Classification Rule Learners*), Redes neuronales (*Neural Networks*), Máquinas de soporte vectorial (*Support Vector Machines*), Bosques Aleatorios (*Random Forest*). Los tres últimos modelos también sirven para regresión.

La problemática que se presenta y que se intentará resolver a partir de los recursos y herramientas enunciados corresponde al área de Marketing: ¿Cómo se puede predecir si el cliente del segmento "Nunca Prestamos" de un Banco tomará o no el préstamo en campaña a través de algoritmos de aprendizaje automático?

Para ello, se comparará la performance de dos algoritmos de clasificación binaria: *Decision Tree* y *Random Forest* (método de ensamble de modelos de *Decision Tree*), utilizando como métrica de performance la exactitud (*Accuracy*) en primer lugar y el área bajo la curva ROC (*AUC ROC*) en segundo lugar. Y se empleará el lenguaje de programación R[3], el cual es libre, gratuito y abierto; y es uno de los lenguajes más empleados en explotación de datos y aplicaciones de *machine learning*.

Cuanto menos heterogéneo sea el universo de partida, mayor será la posibilidad de encontrar relaciones y patrones en común a partir del análisis de su comportamiento histórico y características particulares. Por lo que el foco estará en el segmento de menor tasa de

[3] Creado en 1995 por Ross Ihaka y Robert Gentleman, como descendiente directo del antiguo lenguaje de programación S, R se ha ido fortaleciendo. Escrito en C, Fortran y en sí mismo, el proyecto cuenta actualmente con el apoyo de la R Foundation for Statistical Computing.

conversión a venta: Clientes "No-Anses"[4] del segmento "Nunca Prestamos" de las campañas comerciales de préstamos de pago voluntario.

Las campañas son gestionadas por dos canales: sucursales y centro de llamadas. Tanto el centro de llamadas como las sucursales tienen una capacidad limitada a la cantidad de recursos y horas disponibles de los mismos para la gestión de campañas de cada producto. Y al momento las bases de las campañas superan dicho umbral, lo cual se refleja en un elevado porcentaje de la base sin gestión que ronda el 30-40%.

Cada caso incluido en campaña tiene un costo asociado. Tratándose de casos precalificados por Riesgos, es decir aprobados crediticiamente para obtener un préstamo, conllevan un costo de enriquecimiento de buros. Y a los casos gestionados, se suma el costo de gestión (valor-hora de un recurso). Incluso se podría considerar el costo de oportunidad de esas horas-persona abocadas a cada gestión.

Considerando la capacidad de gestión limitada de ambos canales y el costo asociado a cada caso incluido en campaña, tanto de calificación crediticia (aplicable a todos los casos) como de gestión, resulta indispensable poder realizar una selección criteriosa de casos. Reducir el tamaño de la base, pero mejorar su calidad eligiendo casos con mayor propensión a la toma de préstamos.

Conocer de manera anticipada quienes tomarán el préstamo y quienes no a través de un modelo predictivo, permitiría optimizar la selección ahorrando costos innecesarios, determinar de manera más eficiente la oferta comercial adecuada para ese cliente en función de maximizar la rentabilidad y aumentar la tasa de conversión a venta.

Dada la problemática planteada y la metodología a emplear, el presente trabajo se desarrollará del siguiente modo:

En el capítulo 1, se describirá el contexto tecnológico y de datos al que se enfrentan las organizaciones hoy en día y las implicancias y desafíos que conlleva la aparición del Big Data para la toma de decisiones. Se profundizará en la transformación del área de Marketing y en la reinvención que exigió a sus especialistas.

[4] ANSES (Administración Nacional de la Seguridad Social) es un organismo descentralizado creado en el año 1991, que tiene a su cargo la administración de las prestaciones y los servicios nacionales de la Seguridad Social en la República Argentina.

En el capítulo 2, se verá como los modelos predictivos de aprendizaje automático y la minería de datos vienen a dar sentido a los datos complejos ofreciendo una solución consistente a problemas cotidianos de distintos campos basada en datos reales. Para ello previamente se abordarán conceptos claves como Aprendizaje Automático, Minería de Datos y KDD y su diferenciación

En el capítulo 3, se describirá en detalle el contexto organizacional y coyuntural, la problemática que se plantea y el universo de análisis.

En los capítulos 4 y 5, se propone recorrer las tres primeras fases del total de cinco que componen el proceso metodológico empleado para el desarrollo del modelo, *Knowledge Discovery in Databases* (KDD), a saber:

1. Recopilar e integrar los datos disponibles relativos a la problemática planteada
2. Seleccionar, limpiar y transformar (preprocesamiento) los datos seleccionados.
3. Aplicar técnicas preseleccionadas de minería de datos y aprendizaje automático (construcción del modelo): Arboles de Decisión y Random Forest.

En el sexto y último capítulo, se abarcarán las dos últimas fases del proceso KDD, claves en la extracción de conocimiento y obtención de valor para la compañía a partir de la implementación del modelo en un escenario real:

4. Interpretar y validar la performance del modelo (medidas de evaluación técnica).
5. Evaluación e implementación (obtención del conocimiento, impacto del modelo en resultados).

Capítulo I: Marketing en un Contexto de Big Data

El presente capítulo busca contextualizar al lector. Muestra como los avances tecnológicos abrieron una nueva puerta a la toma de decisiones basada en datos reales (de fuentes tradicionales y digitales, estructurados y multiestructurados) para todas las áreas de una organización. Y al mismo tiempo imponen un desafío a las nuevas y viejas generaciones, a quienes ya conviven desde sus inicios profesionales con estas nuevas herramientas y a quienes deben reinventarse para no quedar fuera del mercado.

Luego se verá la transformación del área de Marketing en este nuevo contexto tecnológico y se brindarán definiciones y terminología del mundo real del Marketing de Big Data, donde toma lugar la problemática que se plantea y, sobre todo, las nuevas y poderosas herramientas para afrontarla. Se abordarán los conceptos de Aprendizaje Automático y Minería de Datos, mostrando su amplia aplicabilidad en distintos campos. Y por último que tipos de algoritmos de Aprendizaje Automático existen y cuáles serán utilizados.

1.1. Avances tecnológicos, desafío para la toma de decisiones

Nuestra capacidad para generar y recopilar datos ha aumentado rápidamente. No sólo todas nuestras transacciones comerciales, científicas y gubernamentales ahora están informatizadas, sino que el uso generalizado de cámaras digitales, herramientas de publicación y códigos de barras también genera datos. Desde el lado de la recolección, las plataformas que escanean texto e imágenes, los sistemas de sensores remotos satelitales y el mundo de banda ancha (World Wide Web) nos han inundado con una enorme cantidad de datos. [5] (Riquelme, Ruiz, & Gilbert, 2006) exponen que la revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, y transmitir. Dada la explosión de disponibilidad de las herramientas informáticas de Hardware y Software para el apoyo de los procesos empresariales, hoy día las empresas generan grandes cantidades de información. Hecho que, sumado al aumento de las capacidades de almacenamiento de datos, hacen que las organizaciones puedan disponer de una gran cantidad y variedad de datos relativos a su actividad diaria (Drucker, 1999, pp. 176-177).

[5] Jiawei Han, Micheline Kamber. 2006. Data Mining: Concepts and Techniques.

Este crecimiento explosivo ha generado una necesidad aún más urgente de nuevas técnicas y herramientas automatizadas que puedan ayudarnos a transformar estos datos en información y conocimiento útiles. Las organizaciones necesitan saber qué está sucediendo ahora, qué es probable que suceda después y qué acciones se deben tomar para obtener los resultados óptimos. Surge una gran oportunidad para la resolución de problemas basada en datos reales y ya no en intuiciones o pálpitos personales del grupo dirigente de la empresa. El análisis de los datos toma cada vez una función vital en el proceso de toma de decisiones. Particularmente, el área de Marketing está experimentando una transformación dramática hacia un mundo de decisiones basadas en datos. Es una de las áreas más creativas de los negocios, y aún necesita poseer ese atributo para tener éxito. Pero la creatividad se juzga cada vez más no sólo en la imaginación humana, sino también en los clics, las conversiones y la sustentación.

Marketing no es, por supuesto, la única área de negocios que experimenta esta transformación. La explotación de datos es una poderosa herramienta que está siendo utilizada para la prevención de fraudes en distintos ámbitos, para la detección temprana de enfermedades, para la gestión eficiente del Turismo, para mejorar la atención al público, para predecir resultados en el Deporte e innumerables campos más. El mundo en general está cada vez más basado en los datos.

Sin embargo, el cambio en Marketing es especialmente llamativo. En poco más de una década, la función ha pasado de enfatizar en imágenes bonitas y frases pegadizas a una que captura, integra y analiza datos de todo tipo.

No es necesario decir que muchos especialistas en marketing, y los gerentes fuera de la función que se relacionan con ellos, no están preparados para esta transformación. Han escuchado el ruido sobre el marketing basado en datos, pero esperan poder retirarse antes de que realmente tengan que cambiar toda su orientación. Pero a menos que tengan más de sesenta años, la jubilación no ayudará mucho. Todos los días, los activos de marketing se digitalizan cada vez más. Todos los días, hay disponible más información sobre las preferencias y comportamientos de los clientes. Todos los días, el costo de oportunidad de no realizar marketing basado en datos se acumula.

A nivel organizacional, algún grupo de personas necesita tomar el liderazgo dentro de las empresas para avanzar hacia una cultura centrada en datos y análisis. Marketing, como la función más afectada por el aumento de datos, y como el recolector y usuario más frecuente de datos de clientes, está en una excelente posición para liderar y liderar con el ejemplo. De acuerdo con Thomas H. Davenport (2013), si Marketing puede centrarse en las promociones

de los clientes, comprendiendo la atribución de los medios digitales a las ventas y segmentando a los mercados, el resto de la organización no puede evitar moverse en la misma dirección basada en datos.

En la mayoría de las organizaciones, no es la única función orientada al cliente. Comparte esa responsabilidad con Ventas, Servicio al cliente y en los últimos años con Experiencia del Cliente. Pero, las organizaciones sentirán la necesidad de aclarar quién es realmente responsable de la información del cliente y de hacer que los datos sean accesibles para otras áreas que lo necesitan. Existe una buena oportunidad para quien tome la iniciativa y demuestre estar a la altura de las circunstancias.

Por supuesto, para hacer eso con éxito, Marketing deberá intensificar su profesionalismo en la gestión de datos. Eso implica disciplina, orientación al proceso y mucho trabajo en la integración de datos. En general, estos no son rasgos que se asocian tradicionalmente con el área, por lo que es necesario realizar algunos cambios. La gestión de datos de Marketing tendrá que adoptar algunos de los enfoques para la higiene de datos (seguridad, copia de seguridad, control de versiones, etc.) que las organizaciones de información y tecnología (IT) han empleado durante décadas. [6]

Lo irónico es que Marketing ha "renegado" durante años de las evaluaciones exhaustivas de IT. En lugar de trabajar conjuntamente en un enfoque profesional para la gestión de datos, a menudo intento evadirlo adquiriendo tecnología y gestionando entornos de datos complejos por su cuenta. Es probable que especialistas de Marketing hayan tenido su base de datos y analizado los sentimientos en las redes sociales de manera más rápida y económica. Sin embargo, este enfoque renegado ha dado lugar a datos fragmentados y aislados de los clientes, así como a algunas ineficiencias en la arquitectura de la tecnología y la gestión de la plataforma.

En el futuro, no se trata de que Marketing reemplace a IT en la gestión profesional de datos, sino que colabore con ellos.

A nivel individual, ahora está claro que los especialistas en Marketing de todos los niveles deben adoptar la tecnología y los datos como elementos clave de sus carteras profesionales. Cada especialista individual necesita replantearse una posición en el continuo que tenga un marketing tradicional, creativo e intuitivo en un extremo (una posición que ya no es sostenible por sí misma), y una gestión de datos digitales sólida en el otro. Si está en el

[6] Lisa Arthur. 2013. Big Data Marketing. Wiley.

extremo orientado a los datos, es posible que no se vea muy diferente de una persona de IT tradicional, aunque se especializará en la administración de datos orientados al cliente.
El trabajo para mantenerse al día con la expansión de IT y el conocimiento de Big Data no se detendrá en el futuro previsible. Nuevos canales para el cliente, nuevas categorías de aplicaciones, nuevos tipos de datos para explotar, y nuevos proveedores y ofertas surgen todo el tiempo.

1.2. La importancia de una Estrategia de Datos para toda la Empresa

La integridad de los datos es una preocupación crítica tanto para el aspecto comercial como financiero de cualquier empresa, ya que garantiza la calidad del servicio brindado al cliente, así como el pago exacto de comisiones a los vendedores, entre otros, y eso también lo convierte en una preocupación crítica para el CFO.
Los datos oportunos y precisos pueden garantizar el resultado exitoso de una venta y ayudar a que un negocio crezca y prospere. Sin embargo, si no se maneja correctamente, los problemas de calidad de los datos pueden derivar en decisiones erradas de impacto irreversible. Pudiendo ser indirectamente una fuente de tensión financiera que ningún CFO quiere experimentar.
Dado el creciente enfoque de la industria en el uso de grandes cantidades de datos para monitorear e identificar formas de mejorar el desempeño de las organizaciones, se ha vuelto indispensable la gobernanza de la información. Consiste en un proceso para monitorear de cerca los datos a lo largo de su ciclo de vida, desde la creación y siguiendo de punta a punta el flujo de la información, a fin de garantizar que sean confiables, oportunos, actualizados, coherentes y precisos.
¿Cuál es exactamente el valor comercial de los datos de calidad? La respuesta es simple: los datos de calidad respaldan el servicio de alta calidad, la investigación rigurosa, la mejor oferta y experiencia para el cliente, el pago preciso, una evaluación de riesgos rentable, respuestas integrales a los auditores y la toma de decisiones estratégicas.
Los costos asociados con la mala calidad de los datos son menos obvios. Estos costos se pueden ocultar, por ejemplo, en los gastos de recursos humanos necesarios para problemas de integridad de datos, mejoras del sistema, corrección de errores, estudios de datos individuales, perdidas de oportunidades de ventas, y en el peor de los casos costos asociados a conflictos legales. Sin considerar el difícilmente mesurable costo de reputación si las consecuencias tuvieran un alcance público.

La función de gobernanza de la información debe trabajar para establecer una atmósfera de cumplimiento e integridad centrada en el cliente. Y concientizar a los empleados acerca del lugar prioritario que ocupa para la organización y de las consecuencias de sus acciones sobre la validez de los datos y el resultado de la compañía.

El CFO debe insistir en un enfoque global y empresarial para lograr y mantener la integridad de los datos. Los pasos clave para este fin incluyen cuantificar financieramente los resultados o beneficios logrados, aprovechar el análisis de datos al tomar decisiones y promover el objetivo general de la integridad de los datos, incluso si eso significa abandonar las estructuras de poder existentes dentro de la organización que pueden haber guiado previamente las decisiones.

Un informe publicado por Oracle en julio de 2012, sobre los desafíos comerciales de Big Data, afirma que se pierde hasta el 14 por ciento de los ingresos de una empresa cuando las empresas no gestionan ni analizan los datos. El informe se basa en una encuesta a 333 ejecutivos de nivel C de empresas estadounidenses y canadienses que abarcan 11 industrias. Y el 97 por ciento de los ejecutivos de nivel C dijeron que aún tenían que hacer cambios para mejorar la optimización de la información.

La gestión de datos ciertamente no es fácil ni económica. Por ejemplo, American Banker informó en agosto de 2012 que la administración de datos a los bancos les cuesta del 7 al 10 por ciento de sus costos operativos.[7]

Los presupuestos para el gobierno de datos diferirán según los objetivos específicos de una organización. Los directores financieros deben apoyar las solicitudes de presupuesto relacionadas con los esfuerzos continuos de gobernanza de la información, así como revisar los informes sobre la calidad e integridad de los datos. Esta es información que debe compartirse regularmente con los equipos de calidad de datos y con el consejo de administración de la organización.

Una vez que una organización cuenta con un programa de gobernanza de la información. Será responsabilidad del CFO asegurar que se realicen progresos.

1.3. Transformación del área de Marketing en un contexto de Big Data

Se abordarán definiciones y terminología útil para el *mundo real del Marketing de Big Data*. Prácticamente cada vez que usamos tecnología en el mundo digitalizado de hoy, ya sea para

[7] Bowen, R., & Smith, A. R. (2014). Developing an enterprisewide data strategy. Healthcare Financial Management, 87.

comunicarnos, comprar, aprender, relajarse o interactuar, dejamos un *rastro de información digital*. Todos esos son datos, a medida que se acumulan con el tiempo en dispositivos y propiedades web, se convierten en *Big Data*. Se convierte en un reflejo de cómo pasamos nuestro tiempo, lo que es importante para nosotros, lo que nos gusta e incluso lo que queremos. Combinando todas estas entradas digitales externas con la información financiera, de marketing, de servicios y demográfica que ya está atrapada dentro de la empresa, se obtendrá realmente *Big Data* y se accederá a una visión de 360 grados del cliente. Los datos generan mejores conocimientos, y esos conocimientos generan mejores interacciones, lo que le permite entregar el mensaje correcto a través del canal correcto en el momento correcto al cliente correcto. Y estas interacciones altamente relevantes y personalizadas son las que permiten ofrecer una experiencia de marca mejoradora en general, que resulta en diferenciación competitiva y mayores ingresos.

Big data, entonces, es una recopilación de datos de fuentes tradicionales y digitales que representa una fuente para el descubrimiento y análisis continuo. Muchas definiciones limitan el *Big Data* a las entradas digitales, como el comportamiento web y las interacciones de redes sociales; sin embargo, no podemos excluir los datos tradicionales derivados de la información de transacciones del producto, los registros financieros y los canales de interacción, como el centro de atención telefónica y el punto de venta. Todo eso también es Big Data, a pesar de que puede verse opacado por el volumen de datos digitales que crece a un ritmo exponencial.

Al definir *Big Data*, también es importante comprender la combinación de *datos no estructurados y multiestructurados* que comprende el volumen de información.

Los *datos no estructurados* provienen de información que no está organizada o fácilmente interpretada por las bases de datos tradicionales o los modelos de datos y, por lo general, contiene mucho texto. Los metadatos, los tweets de Twitter y otras publicaciones en redes sociales son buenos ejemplos de datos no estructurados.

El término *datos multiestructurados* se refiere a una variedad de formatos y tipos de datos y puede derivarse de interacciones entre personas y máquinas, como aplicaciones web o redes sociales. Un gran ejemplo son los datos de registro web, que incluyen una combinación de texto e imágenes visuales junto con datos estructurados como formularios o información transaccional. A medida que la interrupción digital transforma los canales de comunicación e interacción, y los especialistas en marketing mejoran la experiencia del cliente en dispositivos, propiedades web, interacciones cara a cara y plataformas sociales, los *datos multiestructurados* continuarán evolucionando.

(Lisa Arthur, 2013) describe los conceptos de "***volumen, velocidad y variedad***" para enmarcar la discusión de *Big Data*. Algunos añaden otras dos V: veracidad y valor.

El ***volumen*** es la cantidad de datos e incluye datos de fuentes tradicionales y no tradicionales. Piense en los datos transaccionales de un comprador típico de un supermercado, los productos comprados, la frecuencia de cada venta, todo eso es Big Data. Ahora combine esos datos tradicionales con los datos digitales de la página de Facebook de ese comprador. Luego, agregue también las páginas de Facebook de todos los demás compradores. Por sorprendente que parezca, ahora hay el equivalente a 100 terabytes de datos cargados en Facebook todos los días. ¿Cuán grande es? Para darse una idea, 100 terabytes es equivalente al espacio para almacenar 33 millones de canciones. Ese es el volumen de datos cargados en Facebook todos los días ... y esa es solo una plataforma social.

La ***velocidad*** refiere a la velocidad de la información generada y que fluye hacia la empresa. Numerosos analistas han tratado de explicar el alucinante crecimiento exponencial de los datos. Según la International Data Corporation (IDC), habrá 40 zettabytes de datos generados anualmente para 2020. Con 2.8 zettabytes generados en 2012, esto significa que el universo digital se duplicará cada dos años hasta 2020.

La ***variedad*** es el tipo de datos disponibles para las empresas y sus equipos de marketing: datos de sensores, datos de SMS, datos de flujo de clics web. La lista de tipos de datos es larga y continuará creciendo, lo que aumentará la complejidad de esta. Potenciado por el hecho de que Marketing es tanto el usuario final como el generador de *Big Data*.

Finalmente, el ***Marketing de Big Data,*** también conocido como marketing basado en datos, es el proceso de recopilación, análisis y ejecución de los conocimientos derivados de *Big Data* para fomentar la participación del cliente, mejorar los resultados de marketing y medir la rendición de cuentas. Debe combinar toda esta información, proveniente de *datos estructurados y no estructurados* generados por canales tradicionales y digitales, con los datos de la empresa para que el marketing y toda la empresa puedan utilizarla de manera más efectiva.

Capítulo II: Técnicas de aprendizaje automático y sus aplicaciones

En el presente capítulo se abordarán conceptos claves como Aprendizaje Automático, Minería de Datos, KDD y su diferenciación. Y como dichas herramientas vienen a dar sentido a los datos complejos ofreciendo una solución consistente a problemas cotidianos de distintos campos basada en datos reales.

2.1. Aprendizaje Automático, Minería de Datos & KDD

Gran parte de la información disponible hoy en día tiene el potencial de informar la toma de decisiones, si sólo hubiera una forma sistemática de darle sentido a todo.

El campo de estudio interesado en el desarrollo de algoritmos informáticos para transformar datos en acciones inteligentes se conoce como aprendizaje automático (*machine learning*). Este campo se originó en un entorno donde los datos disponibles, los métodos estadísticos y el poder de cómputo evolucionaron rápida y simultáneamente. Esto creó un ciclo de avance que permitió recopilar datos aún más grandes e interesantes.

Un hermano estrechamente relacionado del aprendizaje automático, la minería de datos (*data mining)*, se preocupa por la generación de nuevas ideas a partir de grandes bases de datos. La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados.

Aunque hay un cierto desacuerdo sobre la superposición de los dos campos, un punto potencial de distinción es que el aprendizaje automático tiende a centrarse en realizar una tarea conocida, mientras que la minería de datos se trata de la búsqueda de pepitas de información ocultas. Por ejemplo, puede utilizar el aprendizaje automático para enseñar a un robot a conducir un automóvil, mientras que utilizaría la minería de datos para saber qué tipo de automóviles son los más seguros.[8]

En otras palabras, se puede aplicar el aprendizaje automático a tareas que no impliquen la minería de datos, pero si está utilizando métodos de minería de datos, es casi seguro que esté utilizando el aprendizaje automático.

Desde un punto de vista académico, el término minería de datos es una etapa de un proceso mayor y más antiguo llamado *Knowledge Discovery in Databases* (KDD). Se refiere al

[8] Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información (Han & Kamber, 2001). No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Consta de una serie de cinco fases secuenciales que ampliaremos en los capítulos 4, 5 y 6.

El término "minería de datos" no se acuño sino hasta la década de 1990. Pero su base comprende tres disciplinas científicas entrelazadas: estadística (el estudio numérico de relaciones de datos), inteligencia artificial (inteligencia similar a la humana exhibida por software y/o máquinas) y machine learning (algoritmos que pueden aprender de datos para hacer predicciones). Lo que era antiguo es nuevo otra vez, ya que la minería de datos continúa evolucionando para igualar el ritmo del potencial sin límites del Big Data y poder de cómputo asequible.[9]

2.2. Aplicaciones del Aprendizaje Automático

Lantz, B. (2013). Dada la información de muchas personas, un algoritmo de aprendizaje automático aprende patrones de comportamiento típicos que luego pueden usarse para hacer recomendaciones.

En esencia, el aprendizaje automático está principalmente interesado en dar sentido a los datos complejos. Esta es una misión ampliamente aplicable. Así, por ejemplo, se ha utilizado para predecir los resultados de las elecciones; identificar y filtrar los mensajes de spam del correo electrónico; prever actividad criminal; automatizar las señales de tránsito de acuerdo con las condiciones del camino; producir estimaciones financieras de tormentas y desastres naturales; examinar la rotación de clientes; crear aviones de pilotaje automático y automóviles de conducción automática; identificar individuos con la capacidad de donar; dirigir la publicidad a tipos específicos de consumidores; entre otros.

En el presente trabajo, se utilizará el aprendizaje automático para intentar predecir quienes comprarán un préstamo en campaña de la subpoblación “Nunca Préstamos”.

2.3. Tipos de algoritmos de Aprendizaje Automático

Los algoritmos de aprendizaje automático se pueden dividir en dos grupos principales: “de aprendizaje supervisados” que se utilizan para construir modelos predictivos y “de

[9] SAS Enterprice. “Mineria de datos, qué es y porqué es importante”. https://www.sas.com/es_mx/insights/analytics/data-mining.html

aprendizaje no supervisados" que se utilizan para construir modelos descriptivos. El tipo de algoritmo que se necesite depende de la tarea de aprendizaje que se espera lograr.

Se utiliza un modelo predictivo para tareas que implican, como su nombre lo indica, la predicción de un valor utilizando otros valores en el conjunto de datos. El algoritmo de aprendizaje intenta descubrir y modelar la relación entre la variable objetivo o "target" (a predecir) y las variables independientes o explicativas. A pesar del uso común de la palabra "predicción" para implicar pronósticos, los modelos predictivos no necesariamente tienen que prever eventos futuros. Por ejemplo, un modelo predictivo podría usarse para predecir eventos pasados como la fecha de la concepción de un bebé usando los niveles hormonales de la madre; o podrían usarse en tiempo real para controlar los semáforos durante las horas pico.

La supervisión no se refiere a la participación humana, sino al hecho de que los valores objetivos proporcionan un papel de supervisión, ya que indican al modelo la tarea que necesita aprender. Específicamente, dado un conjunto de datos, el algoritmo de aprendizaje intenta optimizar una función (el modelo) para encontrar la combinación de valores de variables que dan como resultado la salida objetivo.

En función de lo que se busque, deberá emplearse uno de los siguientes 4 enfoques de aprendizaje automático: clasificación, predicción numérica, detección de patrones o agrupación.

La tarea de aprendizaje automático supervisado que se usa con frecuencia para predecir a qué categoría pertenece se conoce como clasificación. Es fácil pensar en posibles usos para un clasificador. Por ejemplo, podría predecir si: un equipo de fútbol ganará o perderá; una persona vivirá más allá de los 100 años; un solicitante incumplirá con un préstamo; un terremoto golpeará el próximo año. O en el presente caso de estudio: predecir si un cliente comprará o no el préstamo que se le ofrece en campaña.

Dentro de los modelos de clasificación, algunos algoritmos de aprendizaje supervisado son: Vecino más cercano (*Nearest Neighbor*), Bayes ingenuo (*naive Bayes*), Árboles de decisión (*Decision Trees*), Estudiantes de reglas de clasificación (*Classification Rule Learners*), Redes neuronales (*Neural Networks*), Máquinas de soporte vectorial (*Support Vector Machines*), Bosques Aleatorios (*Random Forest*). Los tres últimos modelos también sirven para regresión.

En los capítulos 4 a 6, se ampliará sobre el proceso KDD y los modelos empleados en la resolución de la problemática planteada: *Decision Tree* y *Random Forest*.

Capítulo III: Caso de estudio: Contexto y Problemática

En este tercer capítulo se describirá la organización, luego la problemática que se presenta y que se intentará resolver y, por último, el universo de análisis.

3.1. Contexto Organizacional y coyuntural

Más del 65% de las ventas del Banco provienen de las campañas comerciales, y un 80% están explicadas por la venta de préstamos. En campañas comerciales se venden fundamentalmente 3 Tipos de Préstamos (exceptuando el instrumento de gestión de clientes morosos -préstamos de Refinanciación - ya que no responde a fines del área comercial): Préstamos de Pago por Débito Automático (PCD), Préstamos de Pago Voluntario (PVOL) y Préstamos con código (PCC).

Históricamente el banco segmentó su cartera, basándose en el beneficio de ANSES que dio origen a ese cliente más que en el volumen del negocio o en un perfil de rentabilidad (ver gráfico 1.1 en Apéndice).

Cartera de Clientes del Banco

Tipo de Cliente	% participación en la Cartera
Cartera General	43,61%
Jubilados	29,16%
AUH	13,61%
C40	5,98%
Otro Beneficios ANSES	5,16%
Convenio	2,38%
Jurídicas	0,10%

Desde sus orígenes, la mayor fuente de ingresos del banco provino de los clientes del padrón de ANSES. Es decir, aquellas personas que ANSES deriva al Banco para cobrar sus haberes (beneficiarios de jubilaciones, pensiones, AUH, garrafa, plan progresar, etc.). Por lo que el producto ‘estrella’ históricamente fue el préstamo PCD, dirigido a clientes que cobran sus ingresos a través de una cuenta en el Banco, de la cual se debitaría el valor de la cuota del préstamo mensualmente, favoreciendo ratios de mora muy bajos (menores al 5%) y garantizando la rentabilidad del producto.

Por otro lado, están los empleados bajo convenio, gestionados principalmente por vendedores del propio gremio y atados a diferentes acuerdos sindicales, cuyo peso relativo fue reduciéndose a lo largo del tiempo.

Y progresivamente fueron ganando lugar los préstamos de pago voluntario. Aquí el cliente 'target' es mucho más heterogéneo y se encuentra dentro del 43,61% de "Cartera General" (ver Tabla 1). El producto PVOL ha ganado mucho protagonismo en los últimos años en los cuales la rentabilidad del producto PCD se vio afectada por el deterioro del padrón debido a la fuerte competencia y la consecuente fuga de jubilados (pasan a cobrar sus haberes con otras entidades bancarias), y asimismo por la desfavorable coyuntura macroeconómica que redujo notablemente el poder adquisitivo de los jubilados y su porcentaje de disponible para endeudarse (ver Gráfico 1.2 y Figura 3.1 en Apéndice).

Dado este contexto, la importancia de poder predecir la compra de préstamos de pago voluntario fue ganando fuerza en los últimos años y continúa en dicha tendencia.

3.2. Problemática por resolver

Las campañas comerciales son gestionadas por dos canales: sucursales y centro de llamadas. El 30% de la base de las campañas analizadas carece de una gestión cargada en el sistema y mirando sólo el segmento "Nunca Préstamos", dicho porcentaje aumenta al 40% en promedio. El centro de llamadas asegura haber barrido más de una vez la base completa y que el porcentaje que figura "sin contacto" corresponde a casos de teléfonos no válidos, categorizados automáticamente por el discador (al no llegar a un contacto efectivo con el cliente). Al momento no existe un proceso regular y confiable que genere la base de llamados clasificados por el discador. Asimismo, si se mira la historia de gestiones de 10 meses para atrás se observa que el 40% de los casos "sin gestión" lograron un contacto efectivo en el pasado. Los hechos denotan un notable deterioro en la gestión del centro de llamadas. Por otro lado, las sucursales tienen una capacidad limitada a la cantidad de recursos y horas disponibles de los mismos para la gestión de campañas. Y las bases de las campañas superan en muchos casos dicho umbral.

Cada caso incluido en campaña tiene un costo asociado. Tratándose de casos precalificados por Riesgos, es decir aprobados crediticiamente para obtener un préstamo, conllevan un costo de enriquecimiento de buros de $5,14 por caso. Y a los casos gestionados, se suma el costo de gestión de $73,46 por caso (basado en el valor-hora de un recurso). Ver Tablas 2.1 y 2.2 en Apéndice con detalles de costos.

Dada la capacidad de gestión limitada de ambos canales y el costo asociado a cada caso incluido en campaña, tanto de calificación crediticia (aplicable incluso al % sin gestión) como de gestión, resulta indispensable poder realizar una selección criteriosa de casos. Reducir el tamaño de la base, pero mejorar su calidad eligiendo casos con mayor propensión a la toma de préstamos.

Conocer de manera anticipada quienes tomarán el préstamo y quienes no a través de un modelo predictivo, permitiría optimizar la selección ahorrando costos innecesarios y determinar de manera más eficiente la oferta comercial adecuada para cada cliente en función de maximizar la rentabilidad.

Aquellos casos que se sabe con anticipación que no aceptarán la oferta, pueden reasignarse a campañas de otros productos más adecuados a su perfil (quizá productos de ahorro). Incluso, el modelo podría resultar una herramienta útil para la asignación de tasas especiales: resignando margen de rentabilidad para los casos "no tomadores" que cumplen determinadas condiciones, a fin de aumentar las probabilidades de colocación con una oferta más tentadora y competitiva.

3.3. Universo de Análisis

Cuanto menos heterogéneo sea el universo de partida, mayor será la posibilidad de encontrar relaciones y patrones en común a partir del análisis de su comportamiento histórico y características particulares. Por lo que el foco estará en el segmento de menor tasa de conversión a venta: Clientes "*No-Anses*" del segmento "*Nunca Préstamos*" de las campañas comerciales de préstamos de pago voluntario.

"*No-Anses*" porque no pertenecen al padrón que cobra un beneficio a través del banco, por eso se les ofrece un préstamo cuyo cobro no es por débito automático sino pago voluntario.

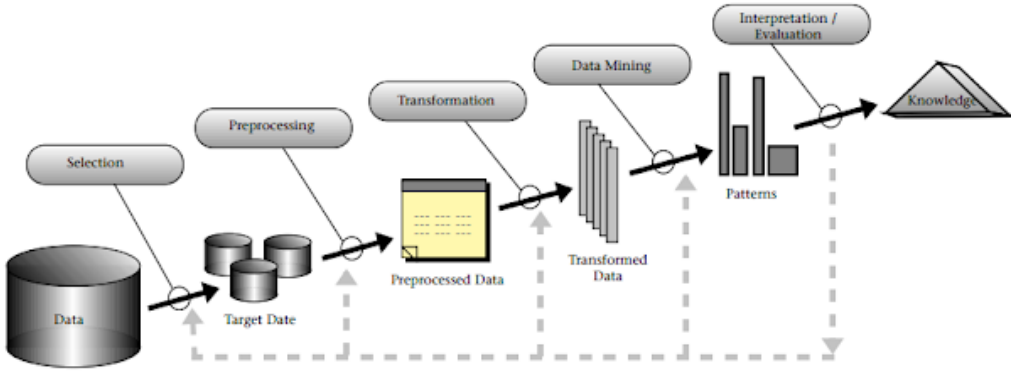
Segmento "*Nunca préstamos*" porque nunca tomaron un préstamo o lo cancelaron hace más de 24 meses.

Se cuenta con información de los participantes de las campañas comerciales de PVOL del 03/01/2019 al 03/09/2019. Siendo las campañas comerciales de periodicidad bimestral, dicho horizonte temporal abarca 4 campañas completas del corriente año (ver tabla 2.3 en el Apéndice).

Capítulo IV: Armado de la Base del Modelo de ML

En el presente capítulo se recorrerán las primeras dos de las cinco fases del proceso *Knowledge Discovery in Databases* (KDD), finalizadas las mismas se habrá construido la base del modelo. La primera fase consiste en la recopilación e integración de los datos disponibles relativos a la problemática planteada. Y la segunda fase consiste en la selección, limpieza y transformación (preprocesamiento) de los datos seleccionados.

Fases del proceso KDD



4.1. Recopilación e Integración de los datos

Los datos utilizados en el presente trabajo se tomaron del ***Datawarehouse del Banco***[10] y del ***Datamart del área de Riesgos,*** ambos son modelos de datos relacional SQL.

Del *Datawarehouse*, se tomó información de varias tablas a saber:

BT_CLIENTES: Tabla con información de todos los clientes del banco (ej. CUIT, Fecha de alta, fecha de nacimiento, código postal, etc.)

BT_GESTIONES_CAMPAÑA: el origen de los datos de esta tabla es el sistema de gestiones de campañas (IBM LOTUS). Dicho sistema es desintegrado del sistema transaccional del banco (COBIS) donde se dan de alta los préstamos al concretarse la venta. LOTUS es un sistema relacional SQL no normalizado[11]. De esta tabla se tomó información de los casos participantes en las 4 campañas consideradas, sus respectivos segmentos, sucursal asignada

[10] Base de datos corporativa que se caracteriza por integrar y depurar información de fuentes distintas, para luego procesarla permitiendo su análisis desde infinitas perspectivas y con grandes velocidades de respuesta.

[11] De acuerdo a Ernesto Chinkes (2001), la normalización es un método que permite estructurar las entidades del modelo de datos de forma tal que sea más eficiente la actualización de los datos, permita ahorrar espacio de almacenamiento y mejorar la recuperación de información de los datos guardados. Para que una entidad se encuentre normalizada debe pasar por distintas "formas normales". Un enfoque podría considerar que una entidad se encuentra normalizada cuando se encuentra en Tercera Forma Normal, lo que implicaría: no contener grupos repetitivos, tener definida su clave primaria, no existir dependencias funcionales parciales respecto de la clave primaria ni existir dependencias funcionales transitivas.

para la gestión, información de si los casos fueron gestionados o no y si tuvieron un contacto efectivo o no.

LK_PRESTAMOS: Tabla que toma la información de altas de préstamos del sistema transaccional (datos reales, no declarativos como las que figuran en Lotus). De aquí se tomó información de los casos que efectivamente convirtieron a venta, a partir de cruces de las tablas *BT_GESTIONES_CAMPAÑA* y *LK_PRESTAMOS*: siempre y cuando el alta del préstamo se haya efectuado en el rango de fechas en el cual el cliente estaba bajo gestión en campaña, la venta se asigna a la misma.

LK_PLAZOSFIJOS: Tabla que contiene información de los plazos fijos al cierre de cada mes. Aquí se tomó información sobre plazos fijos para los casos bajo campaña, si tenían uno o más plazos fijos al inicio de esta y en caso afirmativo, monto total.

CENDEU: Tabla que contiene información pública emitida por el BCRA[12]. De aquí se obtiene con que entidades se encuentra endeudado el cliente, y cuál es su peor situación de mora.

El ***Datamart del área de Riesgos (modelo de datos relacional – SQL),*** en contra de la democratización de la información en la organización, es de acceso y uso exclusivo del sector. Al evaluar una base para una determinada campaña, Riesgos la enriquece con variables de cálculo propio, evaluaciones históricas del motor de riesgos, e información que enriquecen con los Burós (Veraz y P&P). De aquí se toma información sobre las ofertas calculadas para los casos bajo campaña (monto, plazo, cuota), variables de comportamiento de Veraz, nivel de riesgo del cliente.

A partir de la información recopilada de las fuentes enunciadas, se creó un Datamart propio en *SQL Server* llamado "DM_Datamart_Banco", conformado por tablas que contenían información útil para el armado de la base del Modelo. La idea de este Datamart era reunir toda la información necesaria en un único lugar para su análisis e integración en una única base consistente (ver Figura 3.2 en Apéndice).

[12] Central de Deudores: Informe consolidado por clave de identificación fiscal (CUIT, CUIL o CDI) respecto de financiaciones otorgadas por entidades financieras, fideicomisos financieros, entidades no financieras emisoras de tarjetas de crédito / compra y otros proveedores no financieros de créditos y, además, cheques rechazados. https://www.bcra.gob.ar/BCRAyVos/Situacion_Crediticia.asp

4.2. Selección, limpieza y transformación (preprocesamiento)

Luego de recopilar los datos, integrarlos en un mismo Datamart y armar una base preliminar que reuniera todas las variables disponibles de los clientes participantes de campañas, ***se seleccionaron las siguientes 264 variables*** para el desarrollo del modelo:

SUCURSAL_ASIGNADA: sucursal asignada al cliente para su gestión.
SEXO: género femenino o masculino.
EDAD: edad al inicio de la campaña.
ANTIGÜEDAD: antigüedad en el banco al inicio de la campaña.
DIRECCION_CP: código postal de la dirección declarada al alta.
ANTIGUEDAD_VERAZ: cantidad de meses desde que se dio de alta el producto más antiguo. Considera productos activos e inactivos.
ANTIGUEDAD_TC: cantidad de meses desde que se dio de alta la tarjeta de crédito más antigua. Considera productos activos e inactivos.
ANTIGUEDAD_SG: cantidad de meses desde que se dio de alta el préstamo sin garantía más antiguo. Considera productos activos e inactivos.
FIDELIDAD_TC: El índice de fidelidad en Tarjetas de Crédito representa la tendencia en el consumo. Compara los saldos totales de las Tarjetas de Crédito del adherente con respecto al resto del mercado, en los últimos 6 meses. De esta manera, es posible medir tanto la regularidad como el volumen de utilización, mes a mes. El índice toma valores del 0 al 9, siendo 0 cuando no es posible computar la fidelidad, 1 la mínima fidelidad y 9 la máxima fidelidad. La dispersión (ordenada) entre 1 y 9 está determinada por dos factores: la cantidad de meses donde hubo utilización y el saldo total en cada mes.
FIDELIDAD_CC: El índice de fidelidad en Cuentas Corrientes representa la tendencia en el consumo. Compara los saldos al cierre de las Cuentas Corrientes del adherente con respecto al resto del mercado, en los últimos 6 meses. De esta manera, es posible medir tanto la regularidad como el volumen de utilización, mes a mes. El índice toma valores del 0 al 9, siendo 0 cuando no es posible computar la fidelidad, 1 la mínima fidelidad y 9 la máxima fidelidad. La dispersión (ordenada) entre 1 y 9 está determinada por dos factores: la cantidad de meses donde hubo utilización y el saldo al cierre en cada mes.
NIVEL_RIESGO: Nivel de Riesgo del cliente (Alto, Medio, Bajo, Muy Bajo) al inicio de la campaña.

FLAG_MANUAL: 1 cuando requieren aprobación manual para acceder al préstamo (requiere presentación de documentación comprobatoria del Ingreso), 0 cuando tiene aprobación automática del Préstamo.

FLAG_NP_OFERTA: 1 cuando nunca tomó un préstamo con el banco, 0 cuando alguna vez tomó un préstamo con el banco hace más de 24 meses.

MONTO_PF: monto total colocado en plazos fijos al inicio de la campaña.

CANTIDAD_PF: cantidad de plazos fijos vigentes al inicio de la campaña.

FLAG_CLIENTE_PF: 1 si es cliente de plazo fijo, 0 si no lo es.

FLAG_APROB_NOBOOK: 1 cuando es un caso que fue aprobado por el motor de riesgos, pero 'no bookeado' recientemente (últimos dos meses).

FLAG_PNP_RECHAZADO: 1 cuando es un caso que fue rechazado recientemente por el motor de riesgos (últimos dos meses) pero que califica ahora para un préstamo (se le solicita recibo de haberes).

PEOR_SIT_BCRA: peor situación según Informe de CENDEU emitido por el BCRA (disponible al inicio de la campaña). Atraso en el pago: 1. Normal, no supera los 31 días.

2. Riesgo bajo, de más de 31 y hasta 90 días desde el vencimiento. 3. Riesgo medio, de más de 90 y hasta 180 días. 4. Riesgo alto, de más de 180 días hasta un año. 5. Irrecuperable, superiores a un año. 6. Irrecuperable por disposición técnica.

TOTAL_DEUDA_BCRA: Monto total de deuda adquirida con otras entidades registradas en el Informe CENDEU.

FLAG_CENDEU: 1 si figura en el Informe CENDEU (es decir, tiene deuda tomada al menos con otra entidad), 0 si no figura.

CANTIDAD_ENTIDADES: cantidad de entidades con las que figura endeudado según Informe de CENDEU.

FLAG_CENDEU_XXXX: 1 si tiene deuda tomada con la entidad XXXX, 0 si no tiene (cada entidad tiene un código de entidad numérico asignado). La variable de este tipo se construyó para 211 entidades del Informe CENDEU.

USO_CANT*: cantidad de productos activos en el mercado (fuera de la entidad).

USO_SALDO*: acumulado de saldos (uso_tc_saldo + uso_cc_saldo + uso_sg_saldo + uso_gp_saldo + uso_gh_saldo) de productos activos.

USO_LIMITE*: Límite de crédito (uso_tc_limite + uso_cc_limite + uso_sg_limite + uso_gp_limite + uso_gh_limite) de productos activos

USO_EXIGIBLE*: pago mínimo (uso_tc_exigible + uso_cc_exigible + uso_sg_exigible + so_gp_exigible + uso_gh_exigible) de productos activos.

USO_TC_CANT*: cantidad de tarjetas de crédito activas

USO_TC_CANT_A*: cantidad de tarjetas de crédito activas con movimientos en los últimos 6 meses.

USO_TC_SALDO*: sumatoria de saldo total de tarjetas de crédito activas.

USO_TC_LIMITE*: sumatoria del límite de compra de tarjetas de crédito activas.

USO_TC_EXIGIBLE*: sumatoria de pago mínimo de tarjetas de crédito activas.

USO_TC_PROMEDIO3*: promedio de la sumatoria de saldo totales de tarjetas de crédito activas e inactivas en los últimos 3 meses.

USO_TC_PROMEDIO6*: promedio de la sumatoria de saldo totales de tarjetas de crédito activas e inactivas en los últimos 6 meses.

USO_TC_LÍMITE_MAX*: máximo límite de compra de tarjetas de crédito activas.

USO_SG_CANT*: cantidad de préstamos activos sin garantía.

USO_SG_SALDO*: sumatoria de saldos de préstamos activos sin garantía.

USO_SG_LIMITE*: sumatoria de monto acordado de préstamos activos sin garantía

USO_SG_EXIGIBLE*: sumatoria de cuotas + pagos vencidos + intereses de préstamos activos sin garantía.

ACTIVIDAD: cantidad de meses desde la última utilización de los productos activos (en el mercado y en la entidad).

ACTIVIDAD_TC: cantidad de meses desde la última utilización de las tarjetas de crédito activas (en el mercado y en la entidad).

ACTIVIDAD_SG: cantidad de meses desde la última utilización de los préstamos sin garantía (en el mercado y en la entidad).

ACT_OPEN[13]: cantidad de productos abiertos en últimos 6 meses en el mercado.

ACT_OPEN_TC[11]: cantidad de tarjetas de crédito abiertas en últimos 6 meses en el mercado.

ACT_OPEN_SG[11]: cantidad de préstamos sin garantía abiertos en últimos 6 meses en el mercado.

CANT_NO_ENTIDAD_FINAC: cantidad de entidades no financieras con las que figura endeudado en informe CENDEU.

CANT_ENT_FINANC: cantidad de entidades financieras con las que figura endeudado en informe CENDEU.

CANT_ENTIDAD_TC: cantidad de entidades exclusivas de TC con las que figura endeudado en Informe CENDEU.

[13] No considera si los productos fueron dados de baja en el periodo

CANT_ENTIDAD_A: cantidad de entidades de categoría A con las que figura endeudo en Informe CENDEU (a partir de categorización de target de entidades en A, B y C).
CANT_ENTIDAD_B: cantidad de entidades de categoría B con las que figura endeudo en Informe CENDEU (a partir de categorización de target de entidades en A, B y C).
CANT_ENTIDAD_C: cantidad de entidades de categoría C con las que figura endeudo en Informe CENDEU (a partir de categorización de target de entidades en A, B y C).
CANT_MUTUAL: cantidad de mutuales con las que figura endeudado en informe CENDEU.
CANT_COOPERATIVA: cantidad de cooperativas con las que figura endeudado en informe CENDEU.
FLAG_VTA: variable categórica a predecir (1 = venta, 0 = no venta).

* Las variables de "USO" toman para su cálculo toda línea que no pertenezca al grupo de matrices de la entidad/adherente. Siendo:
tc = tarjetas de créditos
cc = descubierto en cuenta corriente
sg = préstamos sin garantía
gp = préstamos con garantía prendaria
gh = préstamos con garantía hipotecaria

Dado que había clientes que participaron en más de una campaña, se aplicaron las siguientes reglas de ***eliminación de duplicados***: si en alguna de las participaciones se logró concretar la venta, se conserva el caso exitoso; si no se concretó la venta en ninguna de las participaciones, se conserva sólo el registro de la campaña más reciente.
Se eliminaron los casos de las bases de campañas que no fueron gestionados, es decir que no tenían una gestión cargada en *Lotus*, ***y los casos que fueron gestionados, pero no lograron un contacto efectivo con el cliente*** (por ejemplo, las resoluciones: teléfono incorrecto, teléfono no valido, no contesta-ocupado, fallecido, hablo con tercero, contestador, etc.). Exceptuando aquellos casos que tuvieron una venta (alta de préstamo en periodo de campaña), y no la cargaron.
También ***se eliminaron los casos cuya resolución cargada en Lotus fue "venta pendiente de cierre" y finalmente la venta no se concretó***, entendiendo que podría tratarse de casos de potencial venta no concretada por errores operativos o mala gestión (aun siendo casos proclives a la toma de un préstamo).

Se procedió a ***imputar los valores perdidos (missing values)*** de las variables de PF (para los no-clientes de PF) y de las variables construidas a partir del Informe CENDEU (casos que no figuran en el Informe emitido por el BCRA y se deduce que no están endeudados con entidades registradas) con el valor "0", indicando la no presencia o existencia para ese registro. Asimismo, las variables construidas FLAG_CLIENTE_PF y FLAG_CENDEU tomarán el valor "0" para estos casos.

Un modelo estimado sobre una base de datos completa donde las clases se encuentran muy desbalanceadas (como en el presente caso: sólo 3,6% de los casos son ventas) tiene menos oportunidad de reconocer diferencias que sobre una base de datos balanceada. Y puede ocurrir que prediga muy bien los casos de 'no venta' (la clase mayoritaria) y no así los de 'venta (la clase minorista), obteniendo un performance "engañosamente" bueno.
Existen diversas **técnicas para balancear la base**. Si por ejemplo quisiéramos obtener una base con una relación de 80%-20% (una relación razonable), algunas técnicas son:
Sobremuestreo (oversampling): se obtiene una muestra mayor, manteniendo los 'no venta' y replicando el 3,6% original de 'venta' hasta llegar a un 20%.
Submuestreo (undersampling): se obtiene una muestra menor, en la que el 3,6% original de 'venta' constituye el 20% de la muestra, reduciendo el número de 'no venta' seleccionados.
Ponderación (weighting): En lugar de desechar individuos (submuestreo) o replicarlos (sobremuestreo), se balancea el conjunto de datos (20-80) asignando un peso en función de si es venta (w1=1) o no (w0<1) y se pondera por esta variable a la hora de estimar el modelo.

En el presente trabajo **se utilizó la técnica de submuestreo bajo criterios semi predefinidos hasta alcanzar una relación 82%-18%**. Se eliminaron los casos de 'no venta' de las campañas ubicadas en los extremos temporales del periodo bajo consideración: campañas 201812 y 201907 (3.523 casos y 10.110 casos, respectivamente). Se eliminaron el 50% de los casos de 'no venta' de cada nivel de riesgo de la campaña 201903 y de la 201905 para mantener los % relativos por nivel de riesgo (2680 y 2255 casos, respectivamente).

Así la base del modelo quedo conformada por 5386 casos. Siendo el 18% casos de venta (FLAG_VTA = 1), 950 casos, y el 82% restante casos de 'no venta' (FLAG_VTA = 0), 4.436 casos.

		Casos	Base Modelo
Limpieza	Dataset_Consolidado	101.112	
	Casos no gestionados o sin contacto efectivo	70.297	30.815
	Casos duplicados	4.350	26.465
	Casos pendientes de cierre 'no venta'	2.511	23.954
Balanceo	Casos 'no venta' de campaña 201812	3.523	20.431
	Casos 'no venta' de campaña 201907	10.110	10.321
	50% casos 'no venta' de campaña 201903	2.680	7.641
	50% casos 'no venta' de campaña 201905	2.255	**5.386**

Finalmente, **creamos un *dataset* de entrenamiento y uno de prueba o validación**. Se necesita un *dataset* de entrenamiento para generar un modelo predictivo, y un dataset de validación para comprobar la eficacia de este modelo para hacer predicciones correctas con "datos no conocidos" (es decir, distintos a los datos con los cuales se entrenó).

Para ello se utilizaron técnicas de muestreo, y se tomó una muestra aleatoria de 1615 observaciones (equivalentes al 30% de la base) como base de validación. Las restantes 3771 observaciones se utilizaron como base de entrenamiento del modelo (70% de la base).

Capítulo V: Arboles de Decisión y Random Forest

En este anteúltimo capítulo se verá la tercer fase clave del proceso *Knowledge Discovery in Databases* (KDD), la aplicación de los algoritmos de aprendizaje automático al dataset trabajado en el capítulo anterior. Es decir, la construcción de los modelos: *Decision Tree* y *Random Forest*.

5.1. Árboles de Decisión

En el campo del aprendizaje automático, hay distintas maneras de obtener árboles de decisión, la utilizada en el presente trabajo es conocida como ***CART: Classification And Regression Trees*** (en español ACR: Árboles de clasificación y Regresión). Esta es una **técnica de aprendizaje supervisado**. Hay una variable objetivo (dependiente) y la meta es obtener una función que permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.

Como el nombre indica, CART es una técnica con la que se pueden obtener árboles de clasificación y de regresión. Se usa clasificación cuando la variable objetivo es discreta, mientras que se usa regresión cuando la variable es continua. Dado que la variable a predecir en el caso bajo estudio (venta = 1, no venta = 0) es una variable discreta, se hará clasificación.

Regresión	Clasificación
Variable dependiente es continua	Variable dependiente es categórica
Valores de los nodos terminales se reducen a la media de las observaciones en esa región.	El valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que han “caído” en esa región.

La implementación particular de CART utilizada es conocida como ***Recursive Partitioning and Regression Trees*** **o** ***RPART***. De allí el nombre del paquete de R empleado.

Este algoritmo divide el espacio de predictores (variables independientes) en regiones distintas y no sobrepuestas, lo que hace es encontrar la variable independiente que mejor separa nuestros datos en grupos que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla (ver gráfico 1.3 en Apéndice). A cada regla corresponde un nodo.

Por ejemplo, la variable objetivo del presente modelo tiene dos niveles, ‘venta’ y ‘no venta’. Encontramos que la variable que mejor separa nuestros datos es ‘Peor situación BCRA’

(descripta más arriba), y la regla resultante es que Peor_Sit_BCRA < 1. Esto quiere decir que los datos para los que esta regla es verdadera tienen más probabilidad de pertenecer al grupo ‘no venta’ (se trata de clientes que no figuran endeudados con ninguna entidad según el informe de CENDEU), que al grupo ‘venta’.

Una vez hecho esto, los datos son separados (particionados) en grupos a partir de la regla obtenida. Después, para cada uno de los grupos resultantes, se repite el mismo proceso. Se busca la variable que mejor separa los datos en grupos, se obtiene una regla, y se separan los datos (en el modelo la segunda regla que mejor separa los datos es Sexo = F). Se repite este procedimiento de manera recursiva hasta que los subgrupos alcancen un tamaño mínimo (preestablecido) o hasta que no se pueda mejorar porque es imposible obtener una mejor separación. Cuando esto ocurre, el algoritmo se detiene. Cuando un grupo no puede ser partido mejor, se le llama nodo terminal u hoja (ver figura 3.3 en el Apéndice).

Una característica muy importante en este algoritmo es que una vez que alguna regla ha sido elegida para separar los datos, ya no es usada de nuevo en los grupos que ha creado. Se buscan reglas distintas que mejoren la separación de los datos.

Además, si después de una partición se han creado dos grupos, A y B. Es posible que para el grupo A, la variable que mejor separa estos datos sea diferente a la que mejor separa los datos en el grupo B. Una vez que los grupos se han separado, al algoritmo “no ve” lo que ocurre entre grupos, estos son independientes entre sí y las reglas que aplican para ellos no afectan en nada a los demás.

El resultado de todo el proceso anterior es una serie de bifurcaciones que tiene la apariencia de un árbol que va creciendo en ramas, de allí el nombre del procedimiento.

En el presente trabajo donde se busca predecir la variable de clasificación binaria venta (1) o no venta del préstamo (0), se arribó al siguiente árbol con las siguientes reglas (ver gráfico 1.4 en Apéndice):

```
n= 3771

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 3771 655 0 (0.826306020 0.173693980)
    2) PEOR_SIT_BCRA< 0.5 1914 110 0 (0.942528736 0.057471264)*
    3) PEOR_SIT_BCRA>=0.5 1857 545 0 (0.706515886 0.293484114)
      6) SEXO=F 1692 381 0 (0.774822695 0.225177305)
       12) EDAD< 63.5 1513 260 0 (0.828155981 0.171844019)
         24) PEOR_SIT_BCRA< 1.5 1442 227 0 (0.842579750 0.157420250)
           48) EDAD< 54.5 1106 138 0 (0.875226040 0.124773960)
             96) FLAG_PNP_RECHAZADO< 0.5 1080 126 0 (0.883333333 0.116666667)*
             97) FLAG_PNP_RECHAZADO>=0.5 26  12 0 (0.538461538 0.461538462)
              194) USO_TC_EXIGIBLE< 1633 12   2 0 (0.833333333 0.166666667)*
              195) USO_TC_EXIGIBLE>=1633 14   4 1 (0.285714286 0.714285714)*
           49) EDAD>=54.5 336  89 0 (0.735119048 0.264880952)
             98) NIVEL_RIESGO=ALTO,BAJO,MUY BAJO 307  68 0 (0.778501629 0.221498371)*
```

```
      99) NIVEL_RIESGO=MEDIO 29   8 1 (0.275862069 0.724137931)*
   25) PEOR_SIT_BCRA>=1.5 71  33 0 (0.535211268 0.464788732)
     50) EDAD< 39.5 14   1 0 (0.928571429 0.071428571)*
     51) EDAD>=39.5 57  25 1 (0.438596491 0.561403509)
      102) ACTIVIDAD_TC< 1.5 13   4 0 (0.692307692 0.307692308)*
      103) ACTIVIDAD_TC>=1.5 44  16 1 (0.363636364 0.636363636)*
 13) EDAD>=63.5 179  58 1 (0.324022346 0.675977654)
   26) ANTIGUEDAD_SG< 31.5 56  24 0 (0.571428571 0.428571429)
     52) TOTAL_DEUDA_BCRA< 67000 45  15 0 (0.666666667 0.333333333)
      104) USO_TC_PROMEDIO6< 1389.5 19   3 0 (0.842105263 0.157894737)*
      105) USO_TC_PROMEDIO6>=1389.5 26  12 0 (0.538461538 0.461538462)
        210) SUCURSAL_ASIGNADA< 70.5 11   2 0 (0.818181818 0.181818182)*
        211) SUCURSAL_ASIGNADA>=70.5 15   5 1 (0.333333333 0.666666667)*
     53) TOTAL_DEUDA_BCRA>=67000 11   2 1 (0.181818182 0.818181818)*
   27) ANTIGUEDAD_SG>=31.5 123  26 1 (0.211382114 0.788617886)
     54) TOTAL_DEUDA_BCRA< 24500 28  13 1 (0.464285714 0.535714286)
      108) SUCURSAL_ASIGNADA< 103.5 15   5 0 (0.666666667 0.333333333)*
      109) SUCURSAL_ASIGNADA>=103.5 13   3 1 (0.230769231 0.769230769)*
     55) TOTAL_DEUDA_BCRA>=24500 95  13 1 (0.136842105 0.863157895)*
7) SEXO=M 165   1 1 (0.006060606 0.993939394)*
```

¿Cómo decide un árbol donde ramificarse?

La construcción de los árboles de decisión estará guiada por minimizar el error de entrenamiento (aun cuando no sea el que nos interese) y se buscarán patrones con poder predictivo que se espera que generalicen bien en datos desconocidos. De esta manera, el objetivo del algoritmo de segmentación recursiva es hacer las variables resultantes en los nodos terminales tan homogéneas como sea posible.

Cuando el árbol es de regresión, se selecciona una “medida de calidad de predicción” (como el error cuadrático medio) y cuando el árbol es de clasificación se utilizan “medidas de impureza” (como Índice Gini, Chi Cuadrado, Ganancia de la información y Reducción en la varianza), como medidas cuantitativas de la homogeneidad.

Los algoritmos para la construcción de árboles de decisión suelen trabajar de manera *top-down*, escogiendo en cada paso la variable que mejor divide el conjunto de elementos. Diferentes algoritmos utilizan diferentes métricas para medir el "mejor". Estos miden generalmente la homogeneidad de la variable de destino dentro de los subconjuntos. Estas métricas se aplican a cada subconjunto candidato, y los valores resultantes se combinan (por ejemplo, un promedio) para proporcionar una medida de la calidad de la división.

En el presente trabajo el algoritmo de ACR utiliza la **impureza de Gini**. Es una medida de cuán a menudo un elemento elegido aleatoriamente del conjunto sería etiquetado incorrectamente si fue etiquetado de manera aleatoria de acuerdo con la distribución de las etiquetas en el subconjunto. La impureza de Gini se puede calcular sumando la probabilidad de cada elemento siendo elegido multiplicado por la probabilidad de un error en la categorización de ese elemento. Alcanza su mínimo (cero) cuando todos los casos del nodo corresponden a una sola categoría de destino.

Para calcular la impureza de Gini de un conjunto de elementos, supongamos i toma valores en {1, 2… m}, y sea f_i la fracción de artículos etiquetados con valor i en el conjunto:

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m}(f_i - {f_i}^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2$$

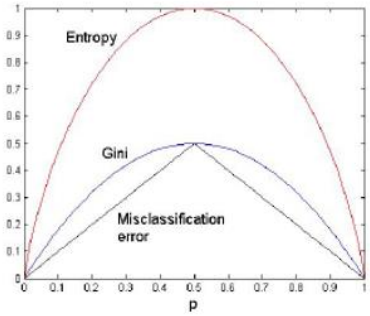
Entonces, dado un nodo t del árbol, Gini(t) mide el grado de "pureza" de t con respecto a las clases. Mayor Gini(t) implica menor pureza. Se buscan divisiones (*splits*) que generen nodos hijos con la mayor pureza posible (o menor impureza posible), es decir cuanto más chico es Gini(t) mejor es.

Gini(t) = 1 – probabilidad de sacar dos registros de la misma clase en el nodo t

Entonces, Gini (t) = probabilidad de NO sacar dos registros de la misma clase del noto t.

The *Gini index* is defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$



En principio el modelo resultante es, con certeza, demasiado complejo, y surge la pregunta como ocurre con todos los procedimientos paso a paso de cuándo detenerse. La segunda etapa del procedimiento consiste en utilizar la validación cruzada para recortar el árbol completo y se explicará más adelante para ambos modelos.

Los principales hiper-parámetros que controlan aspectos del ajuste del Árbol de Decisión (rpart) son: *minplit* (número mínimo de observaciones que deben existir en un nodo para que se intente una división); *minbucket* (número mínimo de observaciones en cualquier nodo terminal. Si solo se especifica uno de minbucket o minsplit, el código establece minsplit en minbucket * 3 o minbucket); *cp* o parámetro de complejidad (no se intenta ninguna división que no disminuya la falta general de ajuste por un factor de cp[14]); *Maxdepth* o máxima profundidad (establece la profundidad máxima de cualquier nodo del árbol final, con el nodo raíz contado como profundidad 0).

[14] Por ejemplo, con la división de Anova, esto significa que el R cuadrado general debe aumentar en cp.

Como el árbol de decisiones es esencialmente un diagrama de flujo, es particularmente apropiado para aplicaciones en las que el mecanismo de clasificación debe ser transparente por razones legales o los resultados deben compartirse para facilitar la toma de decisiones.

Las principales ventajas de este método son: su interpretabilidad (fácil de entender), pues da un conjunto de reglas a partir de las cuales se pueden tomar decisiones; es un algoritmo que no es demandante en poder de cómputo comparado con procedimientos más sofisticados y, a pesar de ello, que tiende a dar buenos resultados de predicción para muchos tipos de datos; útil en exploración de datos (identificar importancia de variables a partir de cientos de variables); menos limpieza de datos (*outliers* y valores faltantes no influencian el modelo a un cierto grado); el tipo de datos no es una restricción, los árboles no requieren escalamiento de atributos o centrado; es un método no paramétrico (i.e., no hay suposición acerca del espacio de distribución y la estructura del clasificador).
Sus principales desventajas son: inestabilidad (puede resultar un tipo de clasificación "débil", en el sentido de que sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo); un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol; sobreajuste, es fácil sobre ajustar los modelos (hacerlos excelentes para clasificar datos que conoce, pero deficientes para datos no conocidos); pérdida de información al categorizar variables continuas.

5.2. Bosques aleatorios (Random Forest)

Random Forest es un modelo tipo ensamblador, que significa grupo. Los métodos tipo ensamblador están formados de un **grupo de modelos predictivos** que permiten alcanzar una mejor precisión y estabilidad del modelo. Estos proveen una mejora significativa a los modelos de árboles de decisión.
Así como todos los modelos, un árbol de decisión también sufre de los problemas de sesgo[15] y varianza[16]. Al construir un árbol pequeño se obtendrá un modelo con baja varianza y alto sesgo. Normalmente, al incrementar la complejidad del modelo, se verá una reducción en el error de predicción debido a un sesgo más bajo en el modelo. En un punto el modelo será muy complejo, se producirá un sobreajuste de este y empezará a sufrir de varianza alta. El modelo óptimo debería mantener un balance entre estos dos tipos de errores. A esto se le

[15] Cuánto en promedio son los valores predichos diferentes de los valores reales
[16] Cuán diferentes serán las predicciones de un modelo en un mismo punto si muestras diferentes se tomarán de la misma población

conoce como “trade-off” (equilibrio) entre errores de sesgo y varianza. El uso de ensambladores es una forma de aplicar este “trade-off” (ver gráfico 1.4 en Apéndice).
Los *Ensambladores* más comunes son *Bagging, Boosting* y *Stacking*. Random Forest es del primer tipo. ***Bagging*** es una técnica usada para reducir la varianza de las predicciones a través de la combinación de los resultados de varios clasificadores, cada uno de ellos modelados con diferentes subconjuntos tomados de la misma población.
Entonces, la idea básica en los modelos “Random Forest”[17] es extraer múltiples muestras aleatorias, con reemplazo, a partir de los datos (este enfoque de muestreo se llama *bootstrap*); usando un subconjunto aleatorio de predictores en cada etapa, ajustar un árbol de clasificación (o regresión) a cada muestra (obteniendo así un "bosque"); combinar las predicciones / clasificaciones de los árboles individuales para obtener predicciones mejoradas. Utiliza la votación para la clasificación (cada árbol vota por una clase) y el promedio para la predicción (promedio de las salidas de todos los árboles), ver figura 3.4 en Apéndice. Así, un grupo de modelos “débiles” se combinan en un modelo robusto. *Random Forest* sirve también como una técnica para reducción de la dimensionalidad.

¿Cómo se construye un modelo *Random Forest*?
Cada árbol se construye así: dado que el número de casos en el conjunto de entrenamiento es N. Una muestra de esos N casos se toma aleatoriamente, pero con reemplazo. Esta muestra será el conjunto de entrenamiento para construir el árbol i. Si existen M variables de entrada, un número m<M se especifica tal que, para cada nodo se seleccionan m variables aleatoriamente de M. La mejor división de estos m atributos es usada para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque. Cada árbol crece hasta su máxima extensión posible y no hay proceso de poda. Nuevas instancias se predicen a partir de la agregación de las predicciones de los x árboles.

Se deben ajustar varios **hiper-parámatros del modelo *Random Forest*** que influyen en su performance. El hiper-parámatros más importante es el número de variables candidatas a seleccionar para evaluar cada ramificación. Sin embargo, existen algunos adicionales que deben considerarse (independientemente de la librería utilizada, los siguientes parámetros deberían estar): *ntree* (número de árboles en el bosque. Se quiere estabilizar el error, pero usar demasiados árboles puede ser innecesariamente ineficiente); *mtry* (número de variables

[17] Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C. Lichtendahl Jr. (2018). Data Mining for Business Analytics. Wiley.

aleatorias como candidatas en cada ramificación); *sampsize* (número de muestras sobre las cuales entrenar; el valor por defecto es 63.25%. Valores más bajos podrían introducir sesgo y reducir el tiempo. Valores más altos podrían incrementar el rendimiento del modelo, pero a riesgo de causar sobreajuste del modelo a la base de entrenamiento. Generalmente se mantiene en el rango 60-80%); *nodesize* (mínimo número de muestras dentro de los nodos terminales. Equilibrio entre sesgo-varianza.); *maxnodes* (máximo número de nodos terminales).

Las principales ventajas de Random Forest son: existen muy pocas suposiciones y por lo tanto la preparación de los datos es mínima; puede manejar hasta miles de variables de entrada e identificar las más significativas (método de reducción de dimensionalidad); una de las salidas del modelo es la importancia de variables; incorpora métodos efectivos para estimar valores faltantes; es posible usarlo como método no supervisado (*clustering*) y detección de *outliers*.

Las principales desventajas de Random Forest son: pérdida de interpretación; bueno para clasificación, no tanto para regresión; las predicciones no son de naturaleza continua; en regresión, no puede predecir más allá del rango de valores del conjunto de entrenamiento; poco control en lo que hace el modelo (modelo caja negra para modeladores estadísticos).

Tuneo de hiper-parámatros con MLR (*GridSearch*) y Validación cruzada (*cross validation*). En el aprendizaje automático, comúnmente se realizan dos tareas al mismo tiempo: validación cruzada y ajuste de hiper-parámatros. La validación cruzada es el proceso de entrenar a los modelos con un conjunto de datos y probarlos con un conjunto diferente para evitar el sobreajuste del modelo (*overfitting*). Y el ajuste de parámetros es el proceso de seleccionar los valores para los parámetros de un modelo que maximizan la precisión del modelo (*tunning*). En este caso, se utilizó la técnica de 'GridSearh', en la cual antes de buscar qué combinación de valores de parámetros produce el modelo más preciso, se debe especificar los diferentes valores candidatos a probar. Y la función del paquete *MLR* probará todas las combinaciones de valores de parámetros posibles y seleccionará el conjunto de parámetros que proporciona el modelo más preciso.

Es así como, bajo los modelos utilizados, los resultados de la aplicación de "tuneo de hiper-parámatros" con *GridSearh* y *Cross-Validation* fueron los siguientes:

Para el Árbol de Decisión: Tune] Result: minsplit=14; minbucket=11; cp=0.001; maxdepth=7

Para el Random Forest: [Tune] Result: ntree=525; mtry=43; nodesize=3

Capítulo VI: Evaluación Técnica de los Modelos y Análisis del impacto de su implementación

En este último capítulo se verán las dos últimas fases de las cinco del proceso *Knowledge Discovery in Databases* (KDD), fundamentales para evaluar el resultado final del camino recorrido. La cuarta fase consiste en la medición de la performance del modelo a partir de las medidas de evaluación técnica preseleccionadas: la exactitud (*Accuracy*) en primer lugar y el área bajo la curva ROC (*AUC ROC*) en segundo lugar. La quinta y última fase consiste en la evaluación e implementación (obtención del conocimiento, resultado del modelo).

6.1. Medición de la performance de los modelos

Las medidas de performance preseleccionadas son Precisión (*Accuracy*) en primer lugar y la curva ROC (AUC-ROC) en segundo lugar.

Matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa a las instancias en la clase real (Referencia), mientras que cada fila representa el número de predicciones de cada clase. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

	Referencia	
Predicción	**No Venta (0)**	**Venta (1)**
No Venta (0)	Verdaderas 'No ventas' (A)	Falsas 'No ventas' (B)
Venta (1)	Falsas 'Ventas' (C)	Verdaderas 'Ventas' (D)

Verdaderas 'no ventas' (A): cantidad de 'no venta' que fueron clasificados correctamente como 'no venta' por el modelo.
Verdaderas 'Ventas' (D): cantidad de 'venta' que fueron clasificados correctamente como 'venta' por el modelo.
Falsas 'no ventas' (B): es la cantidad de 'venta' que fueron clasificados incorrectamente como 'no venta'.
Falsas 'Ventas' (C): es la cantidad de 'no venta' que fueron clasificados incorrectamente como 'venta'.

Exactitud (*Acurracy*): en general, ¿qué porcentaje de la data clasifica correctamente?

$$Exactitud = \frac{(A + D)}{Total}$$

La tasa de precisión general se calcula junto con un intervalo de confianza del 95 por ciento para esta tasa y una prueba unilateral para ver si la precisión es mejor que la "tasa de no información", que se considera el mayor porcentaje de clase en los datos.

Tasa de error (*Misclassification Rate*): en general, ¿qué porcentaje de la data clasifica incorrectamente?

$$Tasa\ de\ error = \frac{(B + C)}{Total}$$

Sensibilidad, exhaustividad, Tasa de verdaderos positivos (*Recall, Sensitivity, True Positive Rate*): cuando la clase es 'no venta', ¿qué porcentaje logra clasificar?

$$Sensibilidad = \frac{A}{(A + C)}$$

Especificidad (*Especificity*), Tasa de verdaderos negativos: cuando la clase es 'venta', ¿qué porcentaje logra clasificar?

$$Especificidad = \frac{D}{(B + D)}$$

Precisión o valor de predicción positivo (Pos Pred Value): cuando predice 'no venta', ¿qué porcentaje clasifica correctamente?

$$Pos\ Pred\ Value = \frac{A}{(A + B)}$$

Valor de predicción negativo (Neg Pred Value): cuando predice 'venta', ¿qué porcentaje clasifica correctamente?

$$Neg\ Pred\ Value = \frac{D}{(C + D)}$$

Exactitud Balanceada (*Balanced Accuracy*):

$$Exactitud\ Balanceada = \frac{(Sensibilidad + Especificidad)}{2}$$

Teniendo en cuenta que la función '*confusionMatrix*' del paquete '*caret*' interpretó como 'Positive' Class: 0 ('no venta'), debajo detalle de los Resultados por modelo:

Decision Tree	no venta	venta	
no venta	4348	472	**4820**
venta	88	478	**566**
	4436	**950**	

Exactitud[18] **=**	0,8960
Tasa de Error =	0,1040
Sensibilidad =	0,9802
Especificidad =	0,5032
Pos Pred Value =	0,9021
Neg Pred Value =	0,8445
Exactitud Balanceada =	0,7417

Random Forest	no venta	venta	
no venta	4422	224	**4646**
venta	14	726	**740**
	4436	**950**	

Exactitud[19] **=**	0,9558
Tasa de Error =	0,0442
Sensibilidad =	0,9968
Especificidad =	0,7642
Pos Pred Value =	0,9518
Neg Pred Value =	0,9811
Exactitud Balanceada =	0,8805

Se observa que el modelo *Random Forest* mejoró no sólo la exactitud respecto al modelo *Decision Tree*, sino todas olas otras métricas de performance calculables a partir de la matriz de confusión. Sobre todo, disminuyó notoriamente la tasa de error y la Especificidad (que en el presente caso representa las clasificaciones correctas sobre la clase 'venta', sumamente importante).

[18] Intervalo de Confianza 95%: (0.8876, 0.9041)
[19] Intervalo de Confianza 95%: (0.95, 0.9611)

Mejoras (RF vs. DT)	
Exactitud =	7%
Tasa de Error =	-58%
Sensibilidad =	2%
Especificidad =	52%
Pos Pred Value =	6%
Neg Pred Value =	16%
Exactitud Balanceada =	19%

El análisis de la **curva ROC**, o simplemente análisis ROC, proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente de (y antes de especificar) el coste de la distribución de las dos clases sobre las que se decide. La curva ROC es también independiente de la distribución de las clases en la población (la prevalencia).

Para dibujar una curva ROC sólo son necesarias las razones de Verdaderos Positivos (VPR) y de falsos positivos (FPR). La VPR mide hasta qué punto un clasificador o prueba diagnóstica es capaz de detectar o clasificar los casos positivos correctamente, de entre todos los casos positivos disponibles durante la prueba. La FPR define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles durante la prueba.

Un espacio ROC se define por FPR y VPR como ejes x e y respectivamente, y representa los intercambios entre verdaderos positivos (en principio, beneficios) y falsos positivos (en principio, costes). Dado que VPR es equivalente a sensibilidad y FPR es igual a 1-especificidad, el gráfico ROC también es conocido como la representación de sensibilidad frente a (1-especificidad). Cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio ROC.

El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). A este punto (0,1) también se le llama una clasificación perfecta. Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación, desde el extremo inferior izquierdo hasta la esquina superior derecha (independientemente de los tipos de base positivas y negativas). Ver gráfico 1.6 en el Apéndice.

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

[0.5]: Es como lanzar una moneda.

[0.5, 0.6): Test malo.

[0.6, 0.75): Test regular.

[0.75, 0.9): Test bueno.

[0.9, 0.97): Test muy bueno.

[0.97, 1): Test excelente.

Para el modelo *Decision Tree* se obtuvo un área bajo la curva (AUC) de 0.816, es decir que según esta medida de performance se obtuvo un modelo bueno (ver gráfico 1.7 en el Apéndice).

```
=== AUCs ===

  Model name Dataset ID Curve type       AUC
1         m1          1        ROC 0.8164147
2         m1          1        PRC 0.6839811
```

Para el modelo *Random Forest* se obtuvo un área bajo la curva (AUC) de 0.977, es decir que según esta medida de performance se obtuvo un modelo excelente (ver gráfico 1.8 en el Apéndice).

```
=== AUCs ===

  Model name Dataset ID Curve type       AUC
1         m1          1        ROC 0.9770249
2         m1          1        PRC 0.9439998
```

Finalmente, si observamos la **importancia de las variables predictoras** en ambos modelos y nos quedamos con las primeras 20 de DT y RF, en ambos modelos coinciden las siguientes 14 variables: PEOR_SIT_BCRA, TOTAL_DEUDA_BCRA, CANTIDAD_ENTIDADES, USO_EXIGIBLE, USO_SALDO, ANTIGUEDAD_SG, ACTIVIDAD_SG, ANTIGUEDAD_TC, SUCURSAL_ASIGNADA, USO_TC_PROMEDIO6, USO_TC_EXIGIBLE, USO_LIMITE, USO_TC_PROMEDIO3, FLAG_CENDEU. Y las tres primeras se encuentran entre las primeras 10 en importancia en ambos modelos.

6.2. Evaluación e implementación: impacto en Resultados del negocio

Se pretende evaluar la efectividad del modelo en términos de su impacto en los resultados del negocio, a partir de su implementación en un escenario anual. Para la simulación de dicho escenario se considerarán: los volúmenes de datos disponibles para 8 meses anualizados, entendiendo que son representativitos de la disponibilidad de casos del segmento "Nunca Préstamos" para la acción de colocación de préstamos de pago voluntario bajo campaña; el Modelo de mejor performance obtenido, *Random Forest.*

La idea es evaluar si el modelo desarrollado, más allá de demostrar una excelente performance a partir de sus medidas de evaluación técnica consideradas (exactitud de 0.9558 y área bajo la curva ROC de 0.977), resulta también una herramienta efectiva para la gestión empresarial eficiente y mejora el rendimiento del capital disponible, mitigando el riesgo operacional y estratégico asumido al realizar acciones de este tipo.

Considerando los costos asociados a cada caso incluido en campaña, de evaluación crediticia por parte de Riesgos de $5.14 por caso - ver Tabla 2.1 en el Apéndice - y de gestión por parte del Centro de llamadas[20] de $73,46 por caso - ver Tabla 2.2 en el Apéndice-, obtenemos un costo total promedio de $78,61 por caso.

Para el costo de gestión, se consideró que un operador de 144 horas mensuales realiza 1400 llamadas al mes y la agencia le cobra al banco $102.850 (Costo + IVA). Y se realiza en promedio una llamada efectiva por caso de campaña (las que no llegan a un contacto efectivo son categorizadas por el discador sin intervención del operador) y el tiempo promedio de la llamada es de es 6' (incluye: tiempo de la llamada 4' + tiempo de carga de la gestión 1.5' + tiempo residual no productivo 0.5').

El préstamo promedio colocado, calculado a partir del monto promedio colocado en la base histórica analizada (4 campañas consideras), es de $31.129 y el spread promedio de un préstamo es de 26 puntos porcentuales (spread promedio a la fecha del análisis).

[20] Al no haber un modelo de asignación de costos que nos permita determinar los valores horas de los oficiales de Sucursal a la gestión de campañas, se utilizará para el presente análisis sólo los costos de gestión disponibles del *call-center* (considerando que el mayor porcentaje es gestionado por dicho canal).

La base de casos anualizada es de 96.762 casos y la ratio de conversión histórico promedio para el segmento "Nunca Préstamos" es del 3.6%, lo que representaría un total de 3473 ventas.

	Base de casos anual
No Venta	93289
Venta	3473
Total *	96762
Tasa conversion (vta)	3,6%

** Dataset_Consolidado (101.112 casos) - casos duplicados (4.350)*

Proyectando la performance del modelo sobre la base de casos anualizada para cada clase, a partir de la matriz de confusión obtenida para RF, podemos observar los siguientes resultados diferenciales bajo un escenario "Sin Modelo" y "Con Modelo":

Resultados del Negocioo Sin Modelo

	Casos camp	Costo	Monto Pmos colocados	Ganancia Bruta (Spread)	Ganancia Neta	Horas call utilizadas
No Venta	55973	$ 4.373.870	$ -	$ -	-$ 4.373.870	5597
Venta	2084	$ 162.849	$ 64.872.836	$ 16.866.937	$ 16.704.088	208
Total	**58057**	**$ 4.536.719**	**$ 64.872.836**	**$ 16.866.937**	**$ 12.330.218**	**5806**

Resultados del Negocioo Con Modelo

	Casos camp	Costo	Monto Pmos colocados	Ganancia Bruta (Spread)	Ganancia Neta	Horas call utilizadas
No Venta	294	$ 22.974	$ -	$ -	-$ 22.974	29
Venta	2654	$ 207.420	$ 82.628.384	$ 21.483.380	$ 21.275.959	265
Total	**2948**	**$ 230.394**	**$ 82.628.384**	**$ 21.483.380**	**$ 21.252.986**	**295**
Variacion monto/cantidad	(55.109)	$ (4.306.325)	$ 17.755.548	$ 4.616.442	**$ 8.922.767**	**(5.511)**
Variacion porcentual (%)	-95%	-95%	27%	27%	**72%**	**-95%**

Se consideró que la actual capacidad de gestión de los canales es del 60% de los casos disponibles, ratio de gestión histórico del segmento "Nunca Préstamos". Es así como bajo el primer escenario "sin modelo" se tomó el 60% de la base, entendiendo que bajo una muestra tomada aleatoriamente (sin ningún criterio establecido) se espera que los pesos de las clases se mantengan. Y bajo el segundo escenario "con modelo", hay una preselección de casos a incluir en campaña determinada por el modelo *Random Forest*. En ambos casos se busca seleccionar una base de campaña que esté dentro de la capacidad de gestión de los canales.
A partir de la aplicación del modelo se obtiene una tasa de conversión 86 p.p. superior a la histórica (2654 ventas sobre una base de 2948 casos implican una tasa de conversión del 90%) y una mejora en la ganancia neta anual de 8,9 millones de pesos (con el modelo la ganancia neta anual es de 21,25 millones de pesos vs. 12,33 millones de pesos en el escenario sin modelo).

Dado que, con la aplicación del modelo se incluyeron menos y más acertados casos, se liberaron 5.511 horas-hombre. Si consideráramos el costo de oportunidad del capital disponible a partir de la aplicación del modelo, la ganancia neta anual podría ser mucho mayor. A modo de referencia se consideró una campaña más rentable como PCD (manteniendo el spread promedio del 26%), donde el monto del prestamos promedio colocado es de $35.000 y la tasa de conversión histórica promedio es del 10%.

Costo de oportunidad de horas liberadas (anual)	
Horas-hombre call liberadas	5.511
Casos gestionables en horas liberadas	55.109
Prestamos PCD colocados esperados	5.511
Monto esperado colocado total	$ 192.885.000
Ganancia Bruta (Spread)	$ 50.150.100
Costo total casos valuados y gestionados	$ 4.306.355
Ganancia Neto (Spread)	**$ 45.843.745**

Entonces, considerando el costo de oportunidad del capital disponible (4,3 millones de pesos para invertir en otra opción), se podría obtener una mejora total en los resultados netos del 344% invirtiendo en campañas de PCD. Lo que implica una ganancia neta total de 54,7 millones de pesos al año mayor a la opción “sin modelo” ($8,9 + $45,8).

Aun considerando inversiones menos redituables que PCD (ya que no siempre es posible conseguir tantos casos como uno quisiera para gestionar) el potencial de generar valor a partir del capital liberado es enorme y muy atractivo para los accionistas.

Conclusión

Se partió de abordar el desafiante contexto tecnológico y de datos al que se enfrentan las organizaciones hoy en día y las implicancias y desafíos que conlleva la aparición del Big Data para las mismas. En este contexto altamente competitivo para la toma de decisiones, se profundizó en la transformación del área de Marketing (principal responsable de generar ingresos a partir de las ventas) y en la reinvención que exigió a sus especialistas.

Bajo un enfoque de gestión de riesgos empresariales (ERM), se explicó el objetivo perseguido con el desarrollo del modelo predictivo de adquisición de préstamos: ofrecer una herramienta útil para mitigar el riesgo empresario que se asume al llevar adelante este tipo de campañas dirigidas al segmento más difícil de convertir ("Nunca Préstamos"), minimizando los efectos del riesgo en el capital y las ganancias de la organización.

Luego se vio como los modelos predictivos de aprendizaje automático y la minería de datos llegaron para dar sentido a los datos complejos ofreciendo una solución consistente a problemas cotidianos de distintos campos basada en datos reales. Se detallaron los distintos algoritmos de aprendizaje supervisado de clasificación existentes, hasta llegar a los modelos empleados en el presente trabajo: Decision Tree y Random Forest.

Reseñando el contexto organizacional y coyuntural, donde toma lugar la problemática abordada se llegó a determinar el universo de análisis.

Seguidamente se recorrieron las fases del proceso metodológico empleado para el desarrollo del modelo, *Knowledge Discovery in Databases* (KDD). Partiendo de la recopilación e integración de los datos disponibles relativos a la problemática planteada. Siguiendo por la selección, limpieza y transformación (preprocesamiento) de los datos seleccionados hasta constituir la base del modelo conformada por 5386 casos y 264 variables (base de entrenamiento y validación) con un 'balanceo' de las clases de 82%-18%. Hasta llegar a la fase de aplicación de las técnicas preseleccionadas de minería de datos y aprendizaje automático (construcción del modelo): *Decision Tree* y *Random Forest*. Casi llegando al final del proceso KDD, la medición del performance técnico a partir de múltiples métricas determinó que el mejor modelo resultó ser *Random Forest* con una exactitud de 0,956 (vs. 0,8960 *Decision Tree*) y un área bajo la curva ROC de 0,977 (vs. 0,8164 *Decision Tree*). Y finalmente, se evaluó la conveniencia de la implementación del mejor modelo obtenido (*Ranom Forest)* a partir de su impacto en los resultados del negocio (ganancia neta obtenida) en un escenario simulado anual, donde se compararon dos escenarios: "Sin Modelo" y "Con Modelo". Se comprobó así que el Modelo produce una mejoría en la ganancia neta del 72,4%

y considerando el costo de oportunidad del capital disponible (liberado por uso eficiente de los recursos al reducirse la base del modelo criteriosamente), se observó una mejoría mayor al 300%. En el presente caso de simulación se consideró su eventual inversión en una campaña de rendimiento conocido como PCD, pero aun considerando inversiones menos redituables (ya que no siempre es posible conseguir tantos casos de PCD como uno quisiera), el potencial de generar valor a partir del capital liberado es enorme y muy atractivo para los accionistas.

Así, se concluye que modelo demuestra ser una herramienta eficiente para la gestión empresarial orientada a crear valor para la compañía y permite mitigar riesgos que atentan contra la certidumbre de las inversiones y el potencial de generar rentabilidad a partir de las mismas, tales como:

El riesgo operacional de error en la presupuestación y planificación comercial, ya que el modelo otorga mayor certidumbre a los accionistas sobre el rendimiento de las inversiones en campañas dirigidas a este segmento de tan difícil conversión, evidenciadas en su exactitud del 98% calculada con un IC del 95%.

El riesgo estratégico de ofrecer el producto adecuado al cliente adecuado[21] (aquel con mayor propensión al consumo de un préstamo) avalado por los valores de predicción positiva y negativa (95,18% y 98,11% respectivamente) calculados en el capítulo 5.

El riesgo estratégico de disponibilidad de capital, liberando 4,3 millones de pesos (reducción del costo), 5.511 horas-hombre que representan 55.109 casos gestionables. Pudiendo destinar esos valiosos recursos a otras acciones que generen mayor valor para la compañía (costo de oportunidad).

El presente modelo fue aplicado en una entidad bancaria, pero podría desarrollarse para compañías de otras industrias y dentro de la misma entidad, podría sentar las bases para desarrollar modelos para otros productos. Llevando a cabo una gestión integral eficiente de los recursos abocados a campañas comerciales a partir de modelos predictivos, y velando siempre por un control periódico de la conveniencia de su implementación y necesidad de posibles ajustes ante cambios de la coyuntura interna y/o externa.

[21] Llamado “customer wants” dentro de los posibles riesgos estratégicos definidos por Casualty Actuarial Society en el “Enterprise Risk Management Committee” de mayo de 2003.

Referencias bibliográficas

Bowen, R., & Smith, A. R. (2014). Developing an enterprisewide data strategy. Healthcare Financial Management, 87.

Drucker, P. F. (1999). Knowledge worker productivity: the biggest challenge. California Management Review, 176–177.

Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C. Lichtendahl Jr. (2018). Data Mining for Business Analytics. Wiley.

Han, J., & Kamber, M. (2001). Data mining concept and technology. Publishing House of Mechanism Industry.

Jiawei Han, Micheline Kamber. 2006. Data Mining: Concepts and Techniques.

Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

Lisa Arthur. 2013. Big Data Marketing. Wiley.

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2018). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

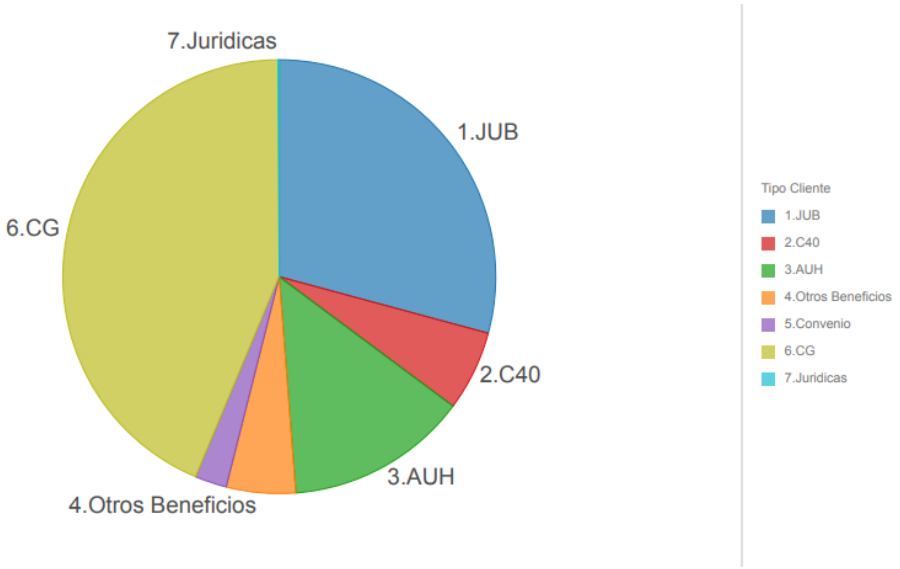
Vijay Kotu, Bala Deshpande, PhD. (2015). Predictive Analytics and Data Mining. Morgan Kaufmann.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Apéndices

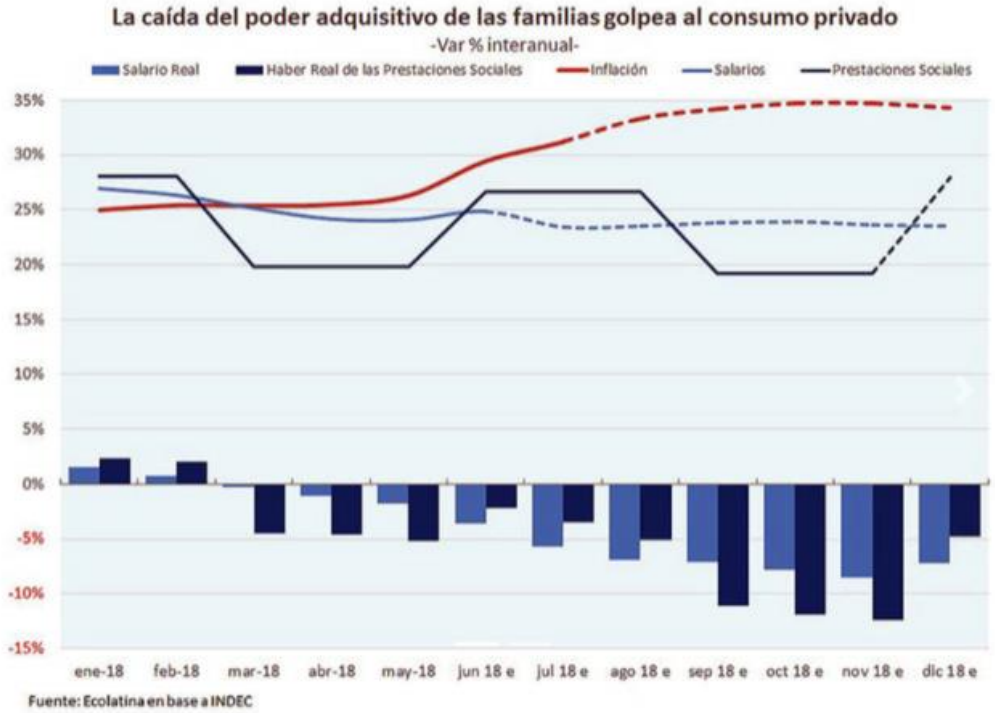
1. Gráficos

Gráfico 1.1 Cartera de Clientes del Banco por Tipo de Cliente



Fuente: Elaboración propia

Gráfico 1.2 Caída del poder adquisitivo de las familias argentinas a lo largo del año 2018



2.
3.
4.
5.

Gráfico 1.3 Lógica de separación en Arboles de Decisión

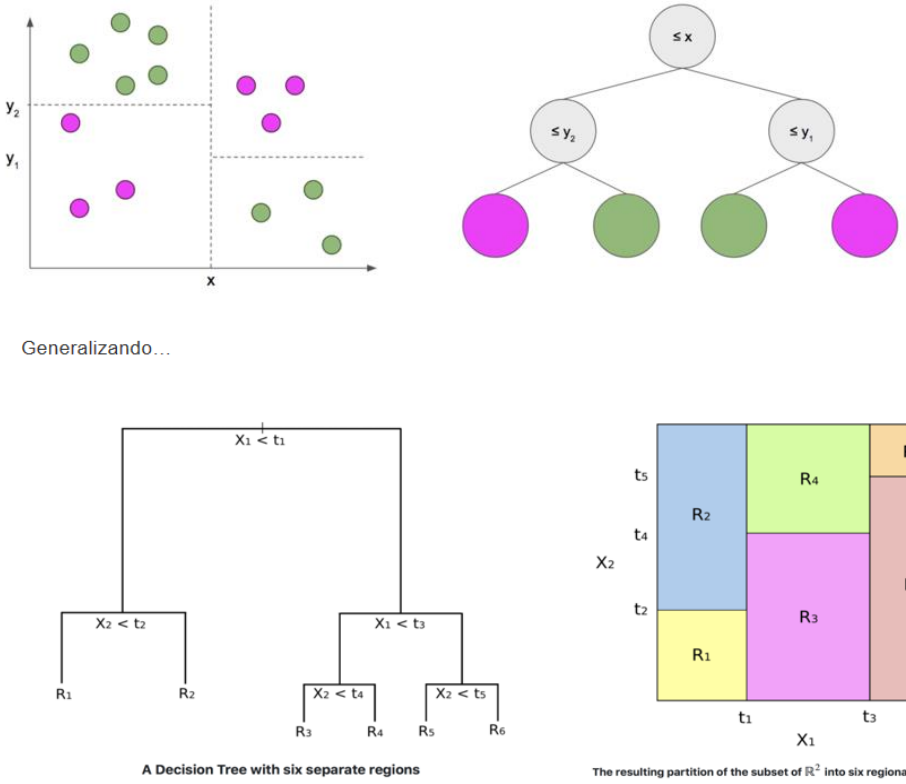


Gráfico 1.4 Relación Sesgo-Varianza en un Modelo

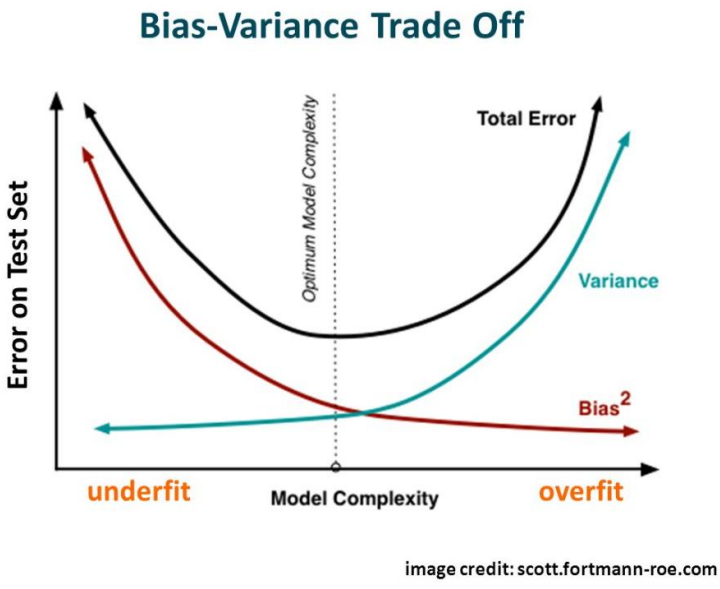


Gráfico 1.5 Árbol de Decisión, Modelo de Adquisición de Préstamos

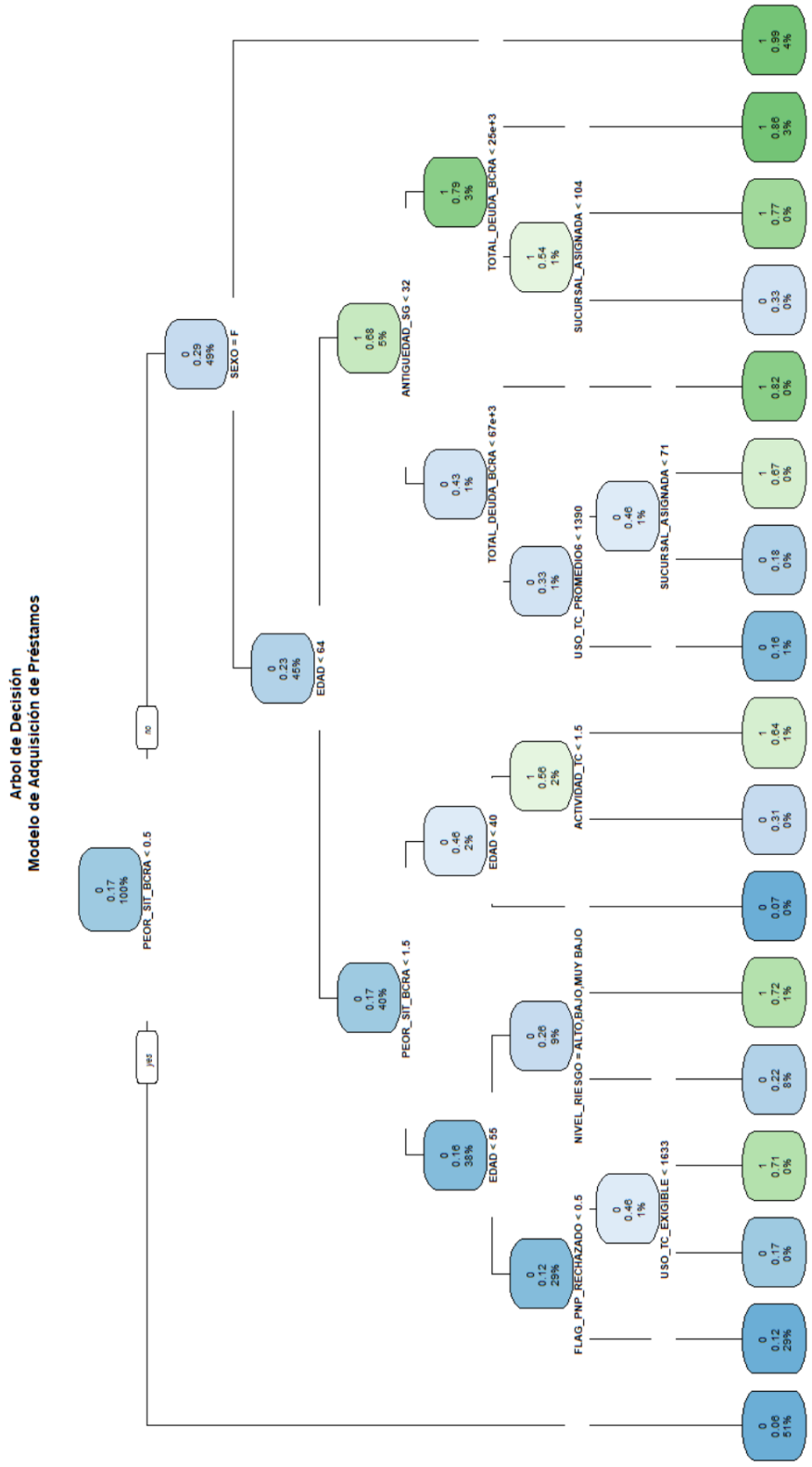


Gráfico 1.6 Curva ROC, diferentes posibles resultados

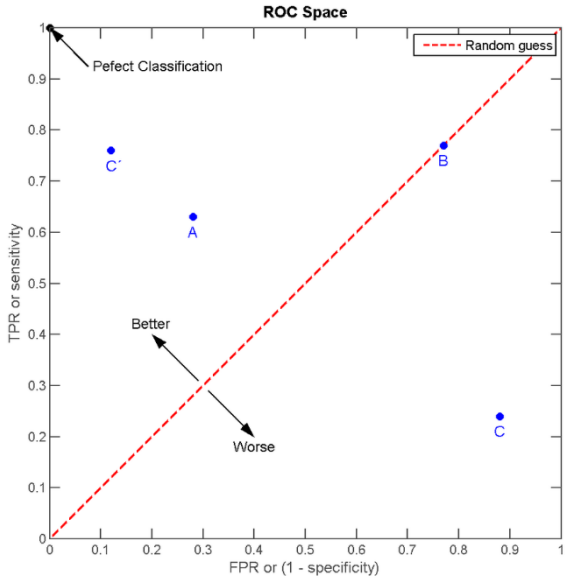


Gráfico 1.7 Curva ROC del modelo Árbol de Decisión para predecir adquisición de préstamos

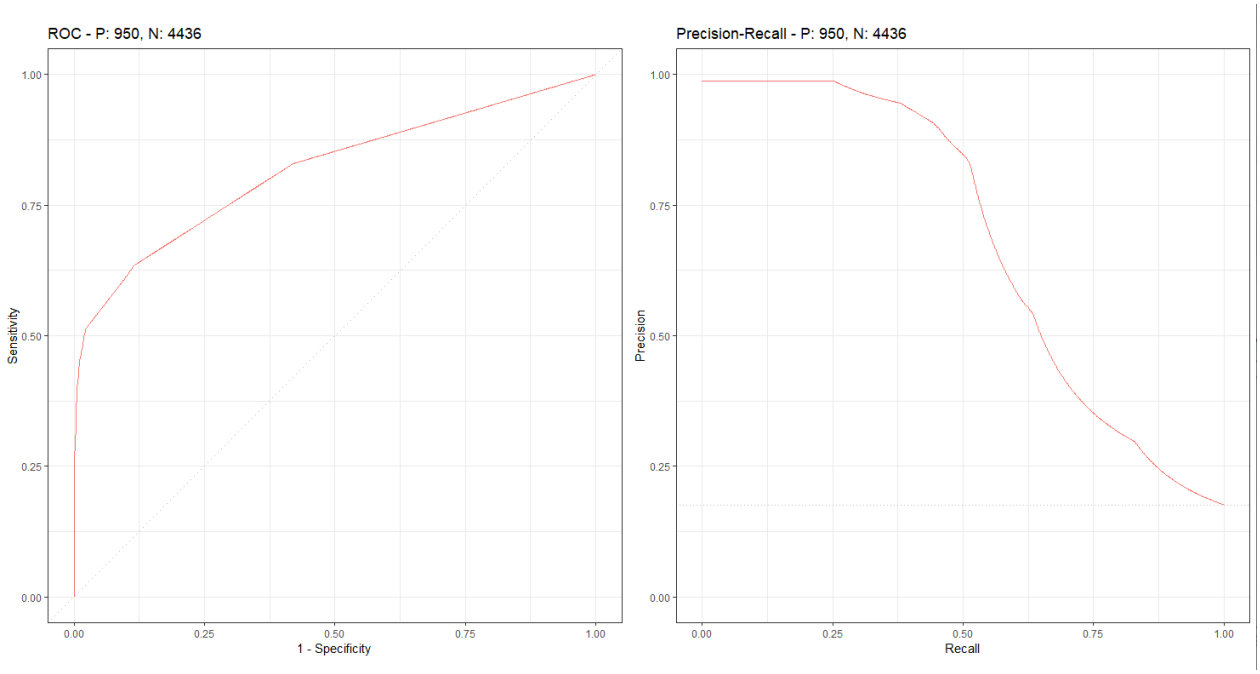
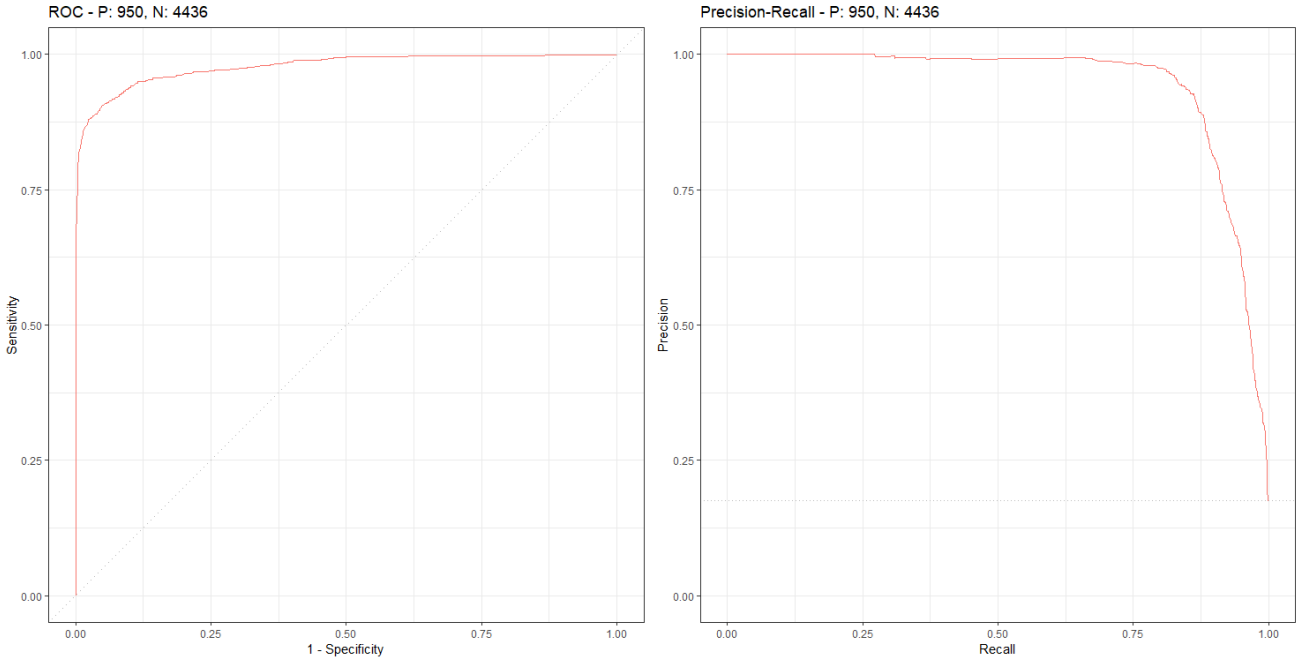


Gráfico 1.8 Curva ROC del modelo Random Forest para predecir adquisición de préstamos



2. Tablas

Tabla 2.1 Costo de calificación de Riegos por caso:

Costo de Buros:

	Costo por Caso	Costo + IVA
Veraz	$ 2,75	$ 3,33
P&P	$ 1,50	$ 1,82
Costo total por caso		**$ 5,14**

Tabla 2.2 Costo de gestión por caso:

Costos de gestion Callcenter:

Periodo Mensual				Por Hora		
Horas	Llamadas	Costo	Costo + IVA	Llamadas x hora	Costo x hora	Costo x llamado
144	1.400	$ 85.000	$ 102.850	10	$ 714	**$ 73,46**

Tabla 2.3 Información básica de las campañas consideradas:

CAMPANA_ID	NOMBRE_CAMP	FECHA_INICIO	FECHA_CIERRE	DURACIÓN (DÍAS)	DURACIÓN (DÍAS LABORABLES)
468	CAMP REGULAR PVOL 201812	03/01/2019	08/03/2019	64	47
491	CAMP REGULAR PVOL 201903	01/03/2019	07/05/2019	67	48
511	CAMP REGULAR PVOL 201905	07/05/2019	04/07/2019	58	43
525	CAMP REGULAR PVOL 201907	04/07/2019	03/09/2019	61	44

3. Figuras

Figura 3.1 Pérdida de poder adquisitivo de las jubilaciones dic 2017 – marzo 2019

Figura 3.2 Datamart *DM_Datamart_Banco* creado en Sql Server, donde integramos todos los datos disponibles

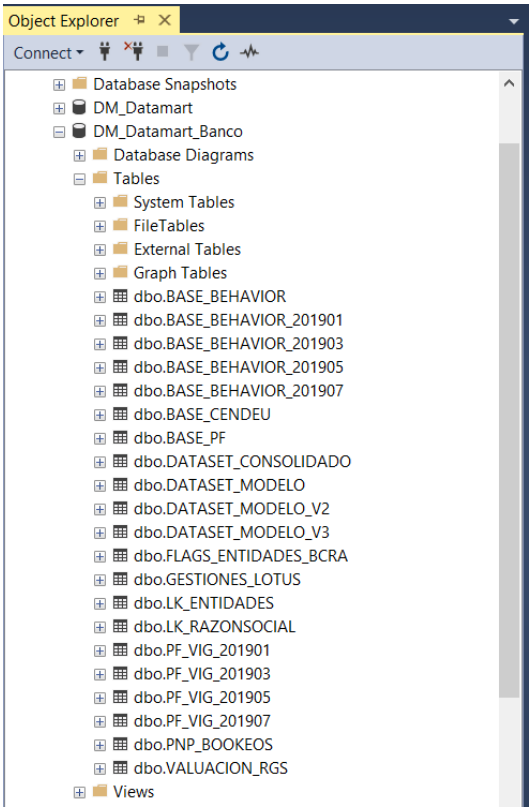


Figura 3.3 Partes de un árbol de Decisión

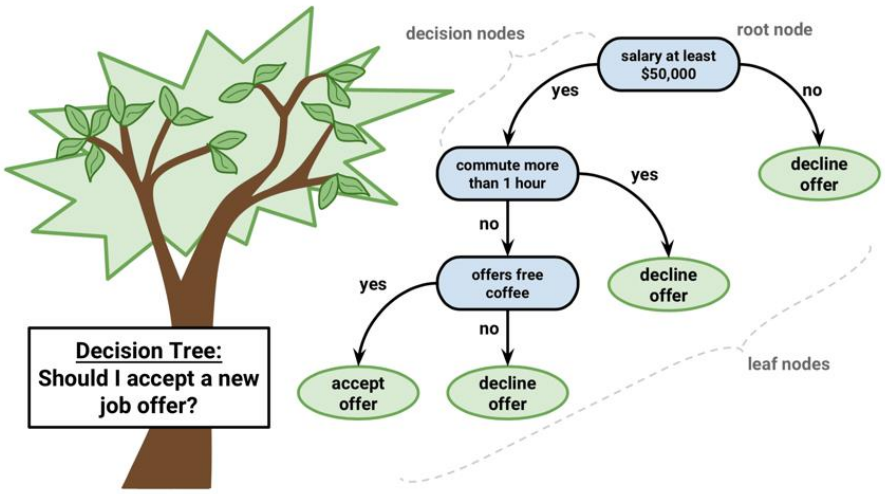
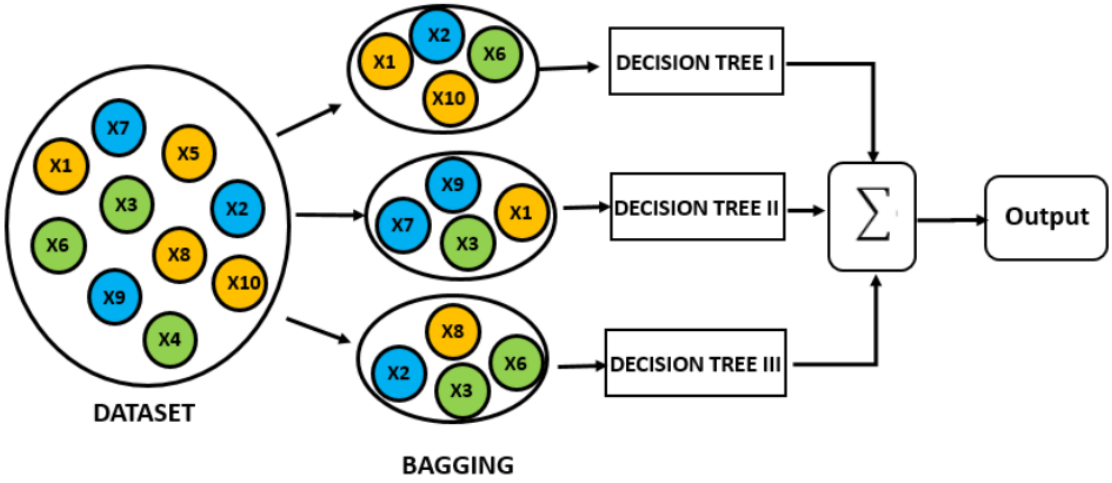


Figura 3.4 Random Forest, técnica de ensamblado de *Bagging*



4. Programación en R:

```
install.packages('tidyverse')
install.packages('mlr')
install.packages('precrec')
install.packages('rpart.plot')
install.packages('ParamHelpers')
install.packages('lattice')
install.packages('caret')

####### Librerías utilizadas:

library(precrec)
library(mlr)
library(tidyverse)
library(rpart)
library(rpart.plot)
library(ParamHelpers)
library(caret)


####### Seteamos ruta de trabajo:
```

```
rm(list=ls())
setwd("C:/Users/danie/Documents /Modelo /Dataset_Modelo ")
getwd()

####### Dataset del modelo:

data_set <- read.csv2("Dataset_Modelo_v5_R.csv", header=TRUE, sep=";")

####### Definimos las variables predictoras y la variable a predecir:

variables = colnames(data_set %>% select(-FLAG_VTA))
predictors = paste(variables, collapse = ' + ')
fmla = paste('FLAG_VTA', predictors, sep = ' ~ ')
data_set$FLAG_VTA = as.factor(data_set$FLAG_VTA)
table(data_set$FLAG_VTA)

####### Dividimos el dataset en base de entrenamiento y validación:

set.seed(1)
val_index <- sample(c(1:nrow(data_set)), 1615) # 30% de la base para validación
train_data <- data_set[-val_index,]
val_data  <- data_set[val_index,]


#**_____________ MODELO 1: DECISION TREE _____________**#

####### A. Tuneo de hiper-parámatros para un Decision Tree con MLR (GridSearch),
usando "cross validation":

trainTask <- makeClassifTask(data = train_data,target = 'FLAG_VTA', positive = "1")

makeatree <- makeLearner("classif.rpart", predict.type = "response")
set_cv <- makeResampleDesc("CV",iters = 5L)

gs <- makeParamSet(
  makeDiscreteParam("minsplit",values =  c(11,12,13,14,15)),       #primero probamos:
values = seq(2,40,2) c(8,9,10,11
  makeDiscreteParam("minbucket", values = c(9,10,11,12)),       #primero probamos:
values = seq(2,12,1) c(9,10,11)
  makeDiscreteParam("cp", values = c(0.0001, 0.001, 0.01 )),
  makeDiscreteParam("maxdepth", values = seq(4,8,1))              #primero probamos:
values = seq(2,20,1) seq(4,6,1)
)

gscontrol <- makeTuneControlGrid()
stune <- tuneParams(learner = makeatree, resampling = set_cv, task = trainTask, par.set =
gs, control = gscontrol, measures = acc)
```

```
stune$x
stune$y

##[Tune] Result: minsplit=14; minbucket=11; cp=0.001; maxdepth=7

####### B. Armamos el modelo con los hiperparámetros del Result_2:

# Entrenamos el modelo:
fit_tree = rpart(formula = fmla, data= train_data, cp= 0.001, minsplit=14, minbucket= 11,
maxdepth=7, method = 'class')

# Predecimos sobre la base de validación:
y_pred_dt = predict(fit_tree, val_data, type = 'prob')[,2]
y_pred_dt_class = predict(fit_tree, val_data, type ="class")
y_test = val_data$FLAG_VTA


# Predecimos sobre el dataset completo:
y_pred_dt_dataset = predict(fit_tree, data_set, type = 'prob')[,2]
y_pred_dt_class_dataset = predict(fit_tree, data_set, type ="class")
y_dataset = data_set$FLAG_VTA


####### C. Importancia de los atributos
fit_tree$variable.importance


####### D. Dibujamos el Árbol de Decisión:
rpart.plot(fit_tree, cex = .50)
rpart.plot(fit_tree, type=2, extra=100,cex = .6) #1
rpart.plot(fit_tree, type=2, extra=100,cex = .7)

rpart.plot(fit_tree,
        type = 2, extra = 106,
        under = FALSE, fallen.leaves = TRUE,
        digits = 2, varlen = 0, faclen = 0, roundint = FALSE,
        cex = .6, tweak = 1,
        clip.facs = FALSE, clip.right.labs = TRUE,
        snip = TRUE,
        box.palette = "auto", shadow.col = 0, main = "Árbol de Decisión\nModelo de
Adquisición de Préstamos")

## Si quisiéramos ver el árbol:

fit_tree
```

```
####### E. Medidas de Performance:

## E.1. Accuracy:

# para base de validación:
accuracy_dt_test = print(mean(y_test == y_pred_dt_class))
# Rdo_accuracy_validation:  0.8879257

# para dataset completo:
accuracy_dt_dataset = print(mean(y_dataset == y_pred_dt_class_dataset))
# Rdo_accuracy_validation: 0.8897141


## Matriz de Confusión:

Matriz_confusion_dt = table(y_pred_dt_class_dataset, y_dataset) ## para el dataset completo
Matriz_confusion_dt
MC_DT = confusionMatrix(y_pred_dt_class_dataset, y_dataset)
MC_DT

## E.2. AUC / Curva ROC y Recall:

# sobre base de validación:
precrec_obj_dt <- evalmod(scores = y_pred_dt, labels = y_test)
autoplot(precrec_obj_dt)
precrec_obj_dt
#Rdo_AUC_validation: 0.8084900

# sobre dataset completo:
precrec_obj_dt_T <- evalmod(scores = y_pred_dt_dataset, labels = y_dataset)
autoplot(precrec_obj_dt_T)
precrec_obj_dt_T
#Rdo_AUC_basecompleta: 0.8164147


#**_____________ MODELO 2: RANDOM FOREST _____________*#

install.packages("ranger")
library(ranger)

####### A. Tuneo de hiper-parámatros para un Random Forest con MLR (GridSearch), usando "cross validation":

getParamSet("classif.randomForest")
```

```
rf <- makeLearner("classif.randomForest", predict.type = "response", par.vals = list(ntree =
200, mtry = 3))
rf$par.vals <- list(
importance = TRUE
 )

rf_param <- makeParamSet(
makeDiscreteParam("ntree",  values = c(500,525,550)),
makeDiscreteParam("mtry", values = c(40,41,42,43,44)),
makeDiscreteParam("nodesize", values  = c(3,4,5,6))
)

rancontrol <- makeTuneControlRandom(maxit = 50L)
set_cv <- makeResampleDesc("CV",iters = 3L)
r_tune <- tuneParams(learner = rf, resampling = set_cv, task = trainTask, par.set =
rf_param, control = rancontrol, measures = acc)

r_tune$x
r_tune$y

## [Tune] Result: ntree=525; mtry=43; nodesize=3


####### B. Armamos el modelo con los hiperparámetros del Result:

# Entrenamos el modelo:
set.seed(1)
fit_randomf = ranger(FLAG_VTA~ ., data = train_data, importance = "impurity",
num.trees =525, mtry = 43 , min.node.size = 3, probability = TRUE,classification = TRUE
)
fit_randomf_class = ranger(FLAG_VTA~ ., data = train_data, importance = "impurity",
num.trees =525, mtry = 43 , min.node.size = 3, probability = FALSE,classification =
TRUE )


# Predecimos sobre la base de validación:
p = predict(fit_randomf, val_data)
p_class = predict(fit_randomf_class, val_data)
y_pred_rf =  p$predictions[,2]
y_pred_rf_class =  p_class$predictions


# Predecimos para el dataset completo:
p_dataset = predict(fit_randomf, data_set)
p_dataset_class = predict(fit_randomf_class, data_set)
y_pred_total_rf =  p_dataset$predictions[,2]
y_pred_total_rf_class = p_dataset_class$predictions
```

```
####### C. Importancia de los atributos
fit_randomf$variable.importance

####### D. Medidas de Performance:

## D.1. Accuracy:

# para base de validación:
accuracy_rf_test = print(mean(y_test == y_pred_rf_class))
# Rdo_accuracy_validation: 0.8897833

# para dataset completo:
accuracy_rf_dataset = print(mean(y_dataset == y_pred_total_rf_class))
# Rdo_accuracy_validation: 0.952098

## Matriz de Confusión:

Matriz_confusion_rf = table(y_pred_total_rf_class, y_dataset) ## para el dataset completo
Matriz_confusion_rf
MC_RF = confusionMatrix(y_pred_total_rf_class, y_dataset)
MC_RF

## D.2. AUC / Curva ROC y Recall:

#para la base de validación:
precrec_obj_rf <- evalmod(scores = y_pred_rf, labels = y_test)
autoplot(precrec_obj_rf)
precrec_obj_rf
#Rdo_AUC_validation: 0.8682357

#para el dataset completo:
precrec_obj_rf_T <- evalmod(scores = y_pred_total_rf, labels = y_dataset)
autoplot(precrec_obj_rf_T)
precrec_obj_rf_T
#Rdo_AUC_basecompleta: 0.9770249
```