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

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

PROYECTO
TRABAJO FINAL DE ESPECIALIZACIÓN

Ventas telefónicas de **microcréditos**.
Diseño de un **modelo de predicción de la probabilidad de éxito** de venta.

AUTOR: MARÍA CLARA ACCAME

TUTOR: PABLO HERRERA

AGOSTO 2019

Contenido

Introducción

El sector de las microfinanzas se caracteriza por atender a un público marginado, no bancarizado y no formalizado, difícil de encuadrar en los estándares de los bancos tradicionales.

La falta de datos formales hace difícil de “encontrar” a los microempresarios, ya sea porque trabajan de manera completamente informal o parte formal y parte informal.

El avance de la tecnología ha permitido visibilizar mucha información nueva, las redes sociales, la compra y venta en portales de internet, los nuevos medios de pago digitales. Estos nuevos datos, generados por los mismos usuarios llevará en un futuro no muy lejano a poder trabajar con datos en este segmento también.

A continuación, se describe la creación de un modelo predictivo para mejorar las ventas de microcréditos desde un call center con los datos básicos obtenidos de bases de contactos y los datos de los mismos llamados.

Es una primera aproximación que puede permitir ordenar los llamados y hacer más eficientes las ventas, aunque seguro es un primer paso que puede seguir desarrollándose con mayor cantidad de datos.

Los microcréditos del mundo a la Provincia de Buenos Aires y el uso eficiente de los datos para la gestión en esta industria.

Mejorar las ventas de microcréditos a través del uso eficiente de los datos.

Provincia Microcréditos S.A. es una empresa perteneciente al Banco Provincia creada en el año 2009 con el fin de atender a los trabajadores independientes y microempresarios de la Provincia de Buenos Aires.

Al trabajar con un sector de la población que desarrolla su actividad económica en parte o en su totalidad de manera informal se presenta como un obstáculo la falta de datos del público objetivo y la necesidad de trabajar uno a uno con los potenciales clientes. Para las acciones de marketing, es decir, para logar que los potenciales clientes se enteren de la oferta crediticia, se depende principalmente del boca a boca, constituyendo el origen del 50% de los créditos nuevos. Para la evaluación de los clientes y determinación del monto a otorgar también se trabaja de forma artesanal dependiendo de un ejecutivo de cuentas que visite el lugar de trabajo del solicitante y determine su capacidad de repago.

A medida que la organización fue creciendo comenzó a construir una base de datos propia que sirve para aprender de la propia experiencia con el objetivo de estandarizar los procedimientos y de aumentar las ventas tomando decisiones a partir de los datos.

La empresa cuenta con el servicio de un call center tercerizado que se contacta con potenciales clientes cuyos datos de contacto se obtienen de listados de terceros con los que se firman convenios de colaboración como mayoristas de alimentos, empresas que fabrican pequeñas maquinarias, concesionarias de motos o autos, municipios (secretarias de producción), cámaras empresarias, etc o listados creados de forma interna por la empresa recolectando las consultas que las personas realizan dejando sus datos para ser contactados, principalmente en formularios web a los que llegan desde distintas redes sociales. El recurso del call center es costoso y limitado y las bases mencionadas tienen un muy bajo nivel de conversión a crédito. Es decir, se llama a un gran número de contactos para lograr la venta de pocos créditos por mes. Actualmente se tiene una conversión de las bases a crédito menor al 1% en las bases de terceros y del 7% en las bases obtenidas desde la web.

La necesidad de aumentar las ventas en la empresa hace que sea necesario buscar fuentes de créditos nuevos por fuera al boca a boca que es lo que actualmente da resultado.

Una dificultad que se presenta en el mercado de las microfinanzas es la gran porción de clientes (60% de la cartera activa) que pertenece a personas no bancarizadas, por lo que no

tienen datos registrados en las fuentes habituales de datos como los bureaus de crédito. A su vez, las microfinanzas son aún incipientes en el mercado argentino, siendo Provincia Microcréditos la empresa con mayor cantidad de créditos otorgados, concentrando más de la mitad del mercado activo de microcréditos a nivel nacional. Esto implica que no haya otras empresas con las que compararse o de las cuáles aprender.

El problema que se pretende resolver es la mejora de las ventas de microcréditos ofrecidos a través del canal de venta call center a partir de bases de datos obtenidas de distintas entidades. Se considera que a través de la aplicación de modelos predictivos se puede mejorar la elección de a qué contactos llamar y en consecuencia aumentar la conversión a crédito de los llamados. Se debe lograr identificar qué variables tenemos previas al llamado al contacto y sirven para determinar la probabilidad de éxito de la venta. Determinar la conveniencia de adquirir a terceros datos adicionales de los contactos a gestionar evaluando el costo y la ganancia probable a obtener. Incrementar la conversión a crédito de los llamados realizados, llamando a menos contactos con mayor probabilidad de éxito. Y determinar la cantidad de horas de call center necesarias para lograr el objetivo de ventas.

La experiencia de los últimos años ha generado datos internos dentro de la empresa que permiten aprender de lo realizado y realizar las tareas de forma más efectiva y eficiente. Cabe destacar que se puede trabajar con los datos gracias que a partir del 2016 la empresa comenzó a desarrollar un data warehouse donde se almacenan todos los datos generados en la operación diaria de la empresa. Los datos transaccionales se almacenan de forma ordenada y normalizada. Los datos generados por el call center, al ser una empresa tercerizada, se obtienen de forma diaria en Excel y con el fin de poder trabajarlos deben ser normalizados y cargados en la base de datos. A su vez, las bases de contactos obtenidas de diferentes entidades también se adquieren en Excel y deben ser normalizadas y enriquecidas con información interna y externa para poder trabajar luego en los modelos de predicción.

En base a lo desarrollado en los párrafos precedentes, intentaremos responder a lo largo de este trabajo la siguiente pregunta; **¿Cómo mejorar la conversión a crédito (ventas) de los llamados realizados por el canal call center para clientes nuevos a través de la aplicación de un modelo de predicción de éxito?**

El objetivo del presente trabajo consiste en proponer lineamientos para mejorar las ventas de los llamados realizados por un call center que presta servicio a la empresa.
Para ello se plantean tres objetivos específicos;

Determinar qué variables se obtienen de los contactos de forma previa al llamado y son útiles para la confección de modelos predictivos.
Predecir qué contactos tienen mayor probabilidad de atender el llamado de forma de poder ordenar los llamados a realizar.
Predecir qué contactos son los más propensos a aceptar las ofertas de crédito.

Suponemos a priori que; a través de la clasificación de los llamados a realizar por el call center se puede mejorar las ventas de microcréditos a clientes nuevos en un 30%.
Existen datos de los contactos a llamar que sirven para clasificarlos con modelos predictivos.
Se pueden ordenar los llamados a realizar por el call center en base a la probabilidad de atender el llamado.
Se pueden seleccionar los llamados a realizar, y en consecuencia mejora efectividad y bajar los costos, en base a la probabilidad de éxito del llamado, considerando como éxito aquellos llamados donde se acepta la oferta de crédito.

Breve historia de las microfinanzas; desde el mundo a la provincia de Buenos Aires.

Las microfinanzas nacen en la década de 1970 en Bangladesh en medio de una hambruna que acechaba al país. En una aldea llamada Jobra a partir de la iniciativa de Muhammad Yunus, quien era profesor en la Universidad en ese momento y trabajaba en dicha aldea con la intención de ayudar a sacar a los aldeanos de la pobreza. Los aldeanos trabajaban con gran esfuerzo en los campos que pertenecían a otros, por jornales muy bajos, y cuando podían dedicar su tiempo a realizar trabajos independientes, en general las mujeres eran las que podían hacerlo, dependían de un prestamista local para abastecerse de la materia prima para fabricar productos que luego debían vender obligatoriamente al prestamista a un precio muy bajo. De esta forma quedaban atrapados en un sistema que los mantenía pobres (Yunus, 2008).

Yunus (2008) refiere en su libro que lo primero que intentó hacer fue buscar ayuda en el banco de la universidad pidiéndoles que presten dinero a los aldeanos pobres para que puedan realizar sus artesanías. Desde los bancos le plantearon que los pobres no era solventes dado que no tenían historial crediticio formal, no tenían avales y eran analfabetos. Luego de varios meses de intentar conseguir ayuda de los bancos Yunus decidió ofrecerse el mismo como avalista de los préstamos, de esta forma, el banco le prestaba dinero a él, y a continuación, él se lo daría a los aldeanos. El banco accedió a esto, y cuando comenzó a prestar dinero el resultado fue sorprendente; los pobres devolvían el préstamo siempre y a tiempo.

En el año 1983 nace así el Banco Grameen, también llamado banco de los pobres, en el marco de una ley creada especialmente para tal propósito.

El Banco Grameen define los programas de microcréditos como pequeños préstamos para gente pobre para proyectos de auto-empleo que generan ingresos, permitiéndoles de cuidar de si mismos y sus familias.

En cuanto al desembarco de las microfinanzas a nuestro país, se remontan al año 2005 donde desde a universidad de Buenos Aires en el marco del Programa de las Naciones Unidas para el desarrollo y siendo el año internacional del microcrédito se publicó un trabajo de investigación donde se realiza una revisión de las microfinanzas en Argentina y se describen los antecedentes a continuación mencionados.

Las microfinanzas en Argentina encuentran sus primeros intentos en 1987 con la fundación Juntos creada por el Banco Provincia; llegó a tener cinco locales en distintos partidos del Gran Buenos Aires, desde los cuales benefició a cinco mil clientes. Esta institución funcionó tres años y luego cerró debido a los efectos inflacionarios.

La inestabilidad económica del país, sumado que las microfinanzas se caracterizan por ser prestamos con un elevado costo de evaluación y de montos bajos lo que la hacen una actividad poco rentable, han marcado la mayoría de los intentos de instituciones y fundaciones en Argentina que han terminado cerrando sus actividades o han subsistido con muy baja escala.

Otras instituciones que han trabajado en el país son;

- Programa Emprender, desde el sector privado, orientado al apoyo crediticio para microempresarios con garantías solidarias.
- El banco mundial de la mujer en Córdoba
- La federación económica en Mendoza (FEM)
- La mutual Balcarce
- La mutual de uniformados en Bahía Blanca
- PROMUDEMI en 1990, un programa impulsado por el Gobierno de la Ciudad de Buenos Aires.
- FIE en 2001, siendo una filial de la empresa que opera en Bolivia.
- Desde la Universidad de Buenos Aires se creo en el año 2000 la asociación civil Avanzar por el Desarrollo Humano. Esta asociación ofrece microcréditos y asistencia técnica a habitantes de villas de emergencia y de áreas carenciadas de la Capital Federal. En su página web publican que en la actualidad llevan otorgados 14.000 microcréditos.

En un panorama más actual Tornero (2018) menciona en un artículo para la revista Forbes que en Argentina hay 51 instituciones microfinancieras y es el octavo país con menos oferentes de microcréditos de la región. A pesar de ello entre junio de 2014 a junio del 2017 el número de instituciones aumentó un 31%, el número de préstamos activos ascendió un 5,22% y la cartera bruta creció un 136%. De todas formas, según la licenciada Marta Bekerman, presidenta de Avanzar, “se estima que la oferta de microcréditos cubre menos de un 5% de la demanda potencial, que se estima en alrededor de 1.300.000 requerimientos”. En nuestro país el ratio crédito sobre PBI es solo del 14% (en Uruguay es del 30%; en Brasil, 66%; en Chile, 107%).

Argentina adhirió a los objetivos del g20 dentro de los cuales el banco mundial en su sitio web desarrolla el relacionado a la inclusión financiera mencionando, entre otras cosas, los siguiente;

“El acceso a servicios financieros facilita la vida cotidiana y ayuda a las familias y las empresas a planificar para todo, desde los objetivos a largo plazo hasta las emergencias imprevistas. Es más probable que, en calidad de titulares de cuentas, las personas usen otros servicios financieros, como créditos y seguros, para iniciar y ampliar negocios, invertir en educación o salud, gestionar riesgos y sortear crisis financieras, todo lo cual puede mejorar su calidad general de vida.”

Finalmente, en cuanto a la empresa dentro de la que se desarrolla el presente estudio; es una empresa fundada en el año 2009 con capitales públicos y privados como una iniciativa del Banco de la Provincia de Buenos Aires. La empresa presta servicios al banco operando dentro de sus mismas sucursales con ejecutivos comerciales propios.
La visión de la empresa plasmada en su sitio web es la de “Ser el líder, referente e innovador de las microfinanzas a nivel nacional e internacional, favoreciendo la inclusión, el desarrollo y calidad de vida del segmento de los microempresarios.” Y su misión; “Promover la igualdad de oportunidades brindando soluciones financieras integrales a los microempresarios de la Provincia de Buenos Aires.” La compañía otorga créditos a personas físicas que desarrollen una actividad económica independiente hace al menos 6 meses, ofreciendo montos de entre 1 y 50 salarios mínimos vitales y móviles, con plazos de entre 6 y 70 meses.

La principal característica del financiamiento otorgado por Provincia Microcréditos responde a los mínimos requisitos que pide a los solicitantes y la forma de evaluación crediticia. Permitiendo el acceso al crédito a personas no bancarizadas y no formalizadas que de otra forma no tendrían la posibilidad de obtener financiamiento y sólo disponen de la atención de prestamistas que cobran altas tasas de interés.

Las microfinanzas en particular; se caracterizan por atender a un sector de la población que trabaja en parte o en su totalidad en la informalidad. Esto genera la falta de información tradicional de los potenciales clientes, tanto para alcanzar el segmento con acciones de marketing tradicional o digital, como para evaluarlos y determinar la capacidad de repago del crédito a otorgar.

Para lo segundo, la evaluación crediticia, la empresa comenzó en el 2009 con el modelo tradicional de microfinanzas, replica del utilizado en Chile desde la década del 90 por Banco Estado Microempresas. Ese modelo consiste en tener un equipo de ejecutivos comerciales en terreno que evalúan a cada potencial cliente en su lugar de trabajo. En la evaluación, que dura aproximadamente una hora y media, se busca determinar la capacidad y voluntad de pago del

cliente, basándose en la documentación formal que tenga, pero también en la realidad del negocio y de la familia del solicitante.

Provincia Microempresas ha logrado en 10 años crecer de forma exponencial superando los 230.000 créditos otorgados en 2019, convirtiéndose en líder de mercado a nivel nacional. A lo largo de este recorrido la empresa fue acumulando datos de sus clientes y de la experiencia, comenzando a utilizar esa información para crear nuevos modelos de evaluación crediticia para el sector. Desde 2017 se trabaja con un equipo de modelos de riesgo y un equipo de inteligencia comercial que tienen por objetivo la creación de modelos comerciales y de riesgo basados en datos para lograr mayor eficiencia en la operación de la empresa. Los modelos de decisión permiten disminuir la subjetividad en la toma de decisiones, realizar las mismas tareas de forma más eficiente e incrementar la cantidad de créditos otorgados sin aumentos en los costos de estructura.

Dentro del país no hay otras instituciones que cuenten con la experiencia y volumen de operaciones de la empresa por lo que se toman como referencia modelos de otros países de la región, principalmente Chile, teniendo en cuenta la particularidad que la población chilena está en un 80% bancarizada y los datos disponibles en el mercado de los potenciales clientes tienen un volumen que no es comparable con la Argentina.

En lo que respecta a las ventas telefónicas dentro de la empresa; la compañía siempre contó con el servicio tercerizado de call center.

En marzo de 2009, un mes después del comienzo de las actividades de la empresa, se empieza a trabajar con el servicio de 4 operadores que se dedicaban exclusivamente a la atención de las llamadas entrantes al 0800. A partir de febrero de 2010 se incorpora a las tareas de los operadores la actividad de llamadas salientes para cobranza de los créditos en mora. A mediados del 2010 se comienzan a realizar llamadas salientes de venta de créditos a contactos nuevos, principalmente provenientes de revistas y guías de comercios locales.

En 2011 se comienzan a realizar llamados a clientes vigentes de la empresa ofreciendo renovar el crédito. En el año 2012 se amplía el servicio a 6 operadores y se finaliza el año con 8 operadores. Ese mismo año se adiciona la tarea de llamados salientes de cobranza preventiva, para clientes que no están en mora pero que en los próximos días debían afrontar el pago de la cuota. En el año 2013 se amplía el servicio a 14 operadores.

En el año 2017 se rediseña la estructura dividiendo el call center entre comercial, con 18 operadores y de cobranzas con 14 operadores.

En Agosto 2018, con la implementación de las campañas de precalificados dentro de la empresa, se deja de realizar llamados de venta a clientes y se avoca exclusivamente al llamado

de contactos nuevos para ofrecer crédito, dichos contactos provenientes de dos fuentes principales; página web y redes sociales en primer lugar, y en segundo de listados provistos por terceros con los que se firma convenio de colaboración, tales como cámaras empresarias, asociaciones, secretarías de producción, etc. En el año 2019 se reduce la cantidad de operadores a 10 para las ventas comerciales.

Planificación e implementación del proyecto

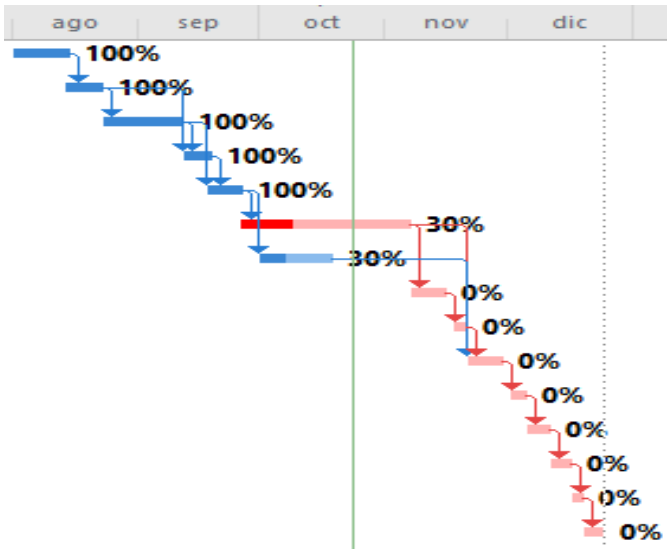
El presente trabajo describe la realización de un proyecto el cual se planificó y ejecuto según metodología tradicional de gestión de proyectos.

Se utilizó el software de Microsoft Project Manager que es un software de administración de proyectos y programas desarrollado y comercializado por Microsoft para asistir a administradores de proyectos en el desarrollo de planes, asignación de recursos a tareas, dar seguimiento al progreso, administrar presupuesto y analizar cargas de trabajo.

A continuación, se observa el cuadro de tareas ejecutadas con la duración en días de las mismas, fecha comienzo y fin, la interdependencia entre tareas, el responsable de cada una y el porcentaje de avance de las mismas.

	Modo de	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras	Nombres de los recursos	% completado
1	✓	Justificación y planteamiento del problema	10 días	jue 1/8/19	mié 14/8/19		Accame María Clara	100%
2	✓	Visación del tutor	7 días	mié 14/8/19	jue 22/8/19	1	Herrera Pablo	100%
3	✓	Marco teórico	14 días	vie 23/8/19	mié 11/9/19	2	Accame María Clara	100%
4	✓	Planteamiento del objetivo general y específicos	5 días	jue 12/9/19	mié 18/9/19	2;3	Accame María Clara	100%
5	✓	Visación del tutor	7 días	mié 18/9/19	jue 26/9/19	3;4	Herrera Pablo	100%
6		Limpieza de la base a trabajar	30 días	jue 26/9/19	mié 6/11/19	5	Accame María Clara	30%
7		Metodología de trabajo	14 días	mar 1/10/19	vie 18/10/19	5	Accame María Clara	30%
8		Diseño del modelo predictivo - entrenamiento	7 días	jue 7/11/19	vie 15/11/19	6	Accame María Clara	0%
9		Visación del tutor	3 días	lun 18/11/19	mié 20/11/19	8	Herrera Pablo	0%
10		Elección del modelo final a trabajar	7 días	jue 21/11/19	vie 29/11/19	9;6;7	Accame María Clara	0%
11		Estadística descriptiva	4 días	lun 2/12/19	jue 5/12/19	10	Accame María Clara	0%
12		Conclusiones	4 días	vie 6/12/19	mié 11/12/19	11	Accame María Clara	0%
13		Corrección del tutor	3 días	jue 12/12/19	lun 16/12/19	12	Herrera Pablo	0%
14		Correcciones finales	3 días	mar 17/12/1	jue 19/12/19	13	Accame María Clara	0%
15		Entrega final	3 días	vie 20/12/19	mar 24/12/19	14	Accame María Clara	0%

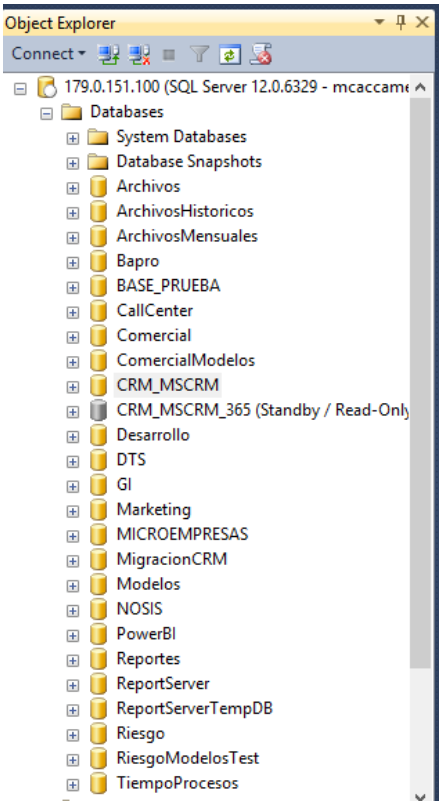
La imagen a continuación muestra gráficamente el cuadro anterior donde se pueden observar en la línea del tiempo las distintas tareas y su interdependencia.



Características de los datos a utiliza para la construcción de los modelos de aprendizaje automático y sus medidas de performance.

Uso de los datos dentro de la organización de microfinanzas y para la construcción de los modelos.

A medida que la organización fue creciendo comenzó a construir una base de datos propia que sirve para aprender de la propia experiencia con el objetivo de estandarizar los procedimientos y de aumentar las ventas tomando decisiones a partir de los datos.



La empresa cuenta con una base de datos administrada desde Microsoft SQL Server Management Studio organizada en distintas bases como se muestra en la imagen a continuación:

Las tablas se almacenan en Microsoft SQL Server Management Studio.

SQL Server es un sistema de gestión de bases de datos relacional lo que le confiere una gran capacidad de gestionar datos, conservando su integridad y su coherencia. (Gabillaud . 2015)

Las funciones que tiene son:

- Almacenar datos
- Verificar restricciones de integridad definidas
- Garantizar la coherencia de los datos que almacena, incluso en caso de error del sistema
- Asegurar las relaciones entre los datos definidas por los usuarios.
- Este motor de base de datos está completamente integrado con Windows lo que le da funcionalidades de seguridad y rendimiento elevados.
- SQL server se ejecuta en forma de servicios de Windows. Según las opciones elegidas se pueden ejecutar más o menos servicios. Los principales son:
- SQL Server: es el servidor de base de datos propiamente dicho. Si no se inicia este servicio no es posible acceder a la información.

- SQL Server Agent: este servicio se encarga de la ejecución de tareas planificadas, la vigilancia de SQL server y el seguimiento de las alertas.
- Microsoft full text search: este servicio se encarga de gestionar la indización de los documentos de tipo de texto almacenados en SQL server y gestionar igualmente la búsqueda de palabras clave.
- El lenguaje natural de SQL Server es Transact SQL. Por tanto, es necesario trasmitir las instrucciones en este lenguaje.

Las bases de datos contienen cierto número de objetos lógicos. Es posible agrupar estos objetos en tres grandes categorías:

- Gestión y almacenamiento de los datos: tablas, tipos de datos, restricciones de integridad, valores por defecto, reglas e índices.
- Acceso a los datos: vistas y procedimientos almacenados.
- Gestión de integridad compleja: triggers (procedimientos almacenados que se ejecutan automáticamente en el momento de la ejecución de una orden SQL que modifique el contenido de una tabla: Insert, Update, y Delete). El trigger está siempre asociado a una tabla y a una instrucción SQL. Permite establecer reglas de integridad complejas entre varias tablas o mantener datos no normalizados.

Los datos con lo que se va a trabajar fueron obtenidos de la base de datos de una empresa de microcréditos que opera en la provincia de Buenos Aires.

Los datos los genera e informa el call center que es una empresa tercerizada que presta servicios a la empresa de créditos.

Se adicionaron además variables de distintas fuentes como el Banco Central de la República Argentina y de la AFIP y variables creadas como la edad, el año de nacimiento que se infieren del número de DNI de la persona a llamar, la hora del llamado, y la franja horaria del llamado.

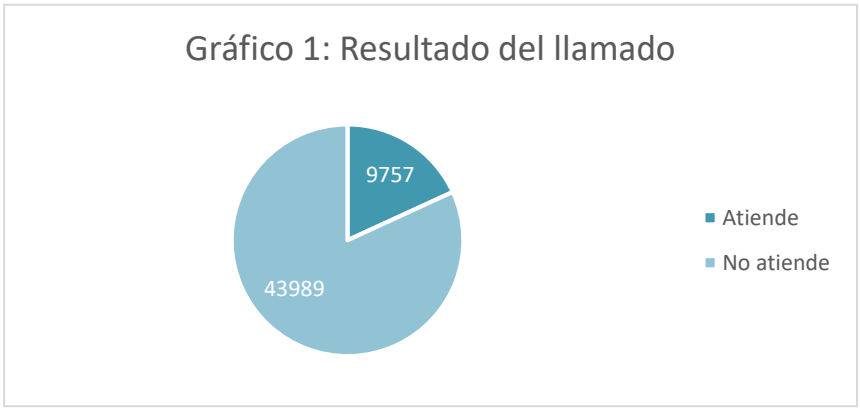
Al finalizar con la creación de nuevas columnas se eliminan de la base a trabajar las columnas que identifican a cada persona; nombre y apellido, dni y cuil, de forma de trabajar con una base anonimizada. En el caso de requerir trackear un llamado en particular o realizar una auditoria la base contiene la columna IDINTERACCION que nos permite identificar cada llamado en particular.

La base contiene 53.746 filas con datos relacionados a cada llamado.

La base contiene 420.597 valores vacíos y 1.626.333 datos no nulos.

Se detalla la totalidad de las variables en el anexo del presente trabajo.

La variable por predecir en primera instancia es la de respuesta al llamado que presenta el comportamiento del gráfico de torta siguiente (gráfico 1):

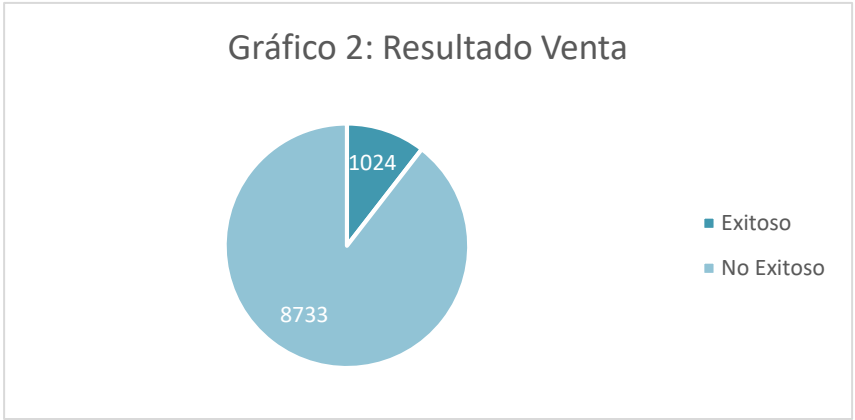


Para el primer modelo, donde vamos a predecir la probabilidad de que el llamado sea atendido o no lo sea nos quedaremos sólo con las variables que hacen a información que tenemos previo a realizar el llamado.

Es decir, no se incorporan en el modelo datos que obtenemos a través del llamado.

Para el segundo modelo, donde predeciremos la probabilidad de venta exitosa usaremos todas las variables contenidas en la tabla a excepción del resultado de venta que es la variable por predecir.

La variable por predecir en segunda instancia es la de “tipos tipificación” que hace referencia a si la venta es exitosa o no y que presenta el comportamiento del gráfico de torta siguiente (gráfico 2):



Modelos predictivos de aprendizaje automático.

Un modelo predictivo es un modelo de datos, basado en estadísticas inferenciales, que se utiliza para predecir un comportamiento futuro o desconocido basándose en la información que se tiene de comportamientos pasados. Se busca determinar la probabilidad de un resultado dada una cantidad de datos de entrada. Los modelos pueden utilizar una o más variables para determinar la probabilidad de un conjunto de datos.

Algunos conceptos básicos que deben tenerse en cuenta y que se definen a continuación son; conjunto de datos, modelo, aprendizaje.

Conjunto de datos: es la información utilizada para construir el modelo. Se puede dividir en dos tipos; conjunto de entrenamiento y conjunto de prueba. El conjunto de entrenamiento se utiliza para determinar los parámetros del clasificador y la otra parte, llamada conjunto de prueba; se utiliza para estimar el error de generalización. El conjunto de entrenamiento suele dividirse a su vez en subconjuntos que sirven para ajustar el modelo y evitar el sobreajuste.

Modelo o clasificador: es una conexión entre las variables que son dadas y las que se van a predecir. Usualmente las variables que se van a predecir son denominadas dependientes y las restantes independientes.

Aprendizaje: es cualquier procedimiento utilizado para construir el modelo a partir del conjunto de datos de entrenamiento. Es el proceso mediante el cual un sistema mejora y adquiere destreza en la ejecución de sus tareas y tiene la capacidad de poseer inferencia inductiva sobre estas. El aprendizaje inductivo es en el que se crean modelos a partir de generalizar ejemplos. Se buscan patrones comunes que expliquen ejemplos y que se puedan aplicar luego sobre información nueva.

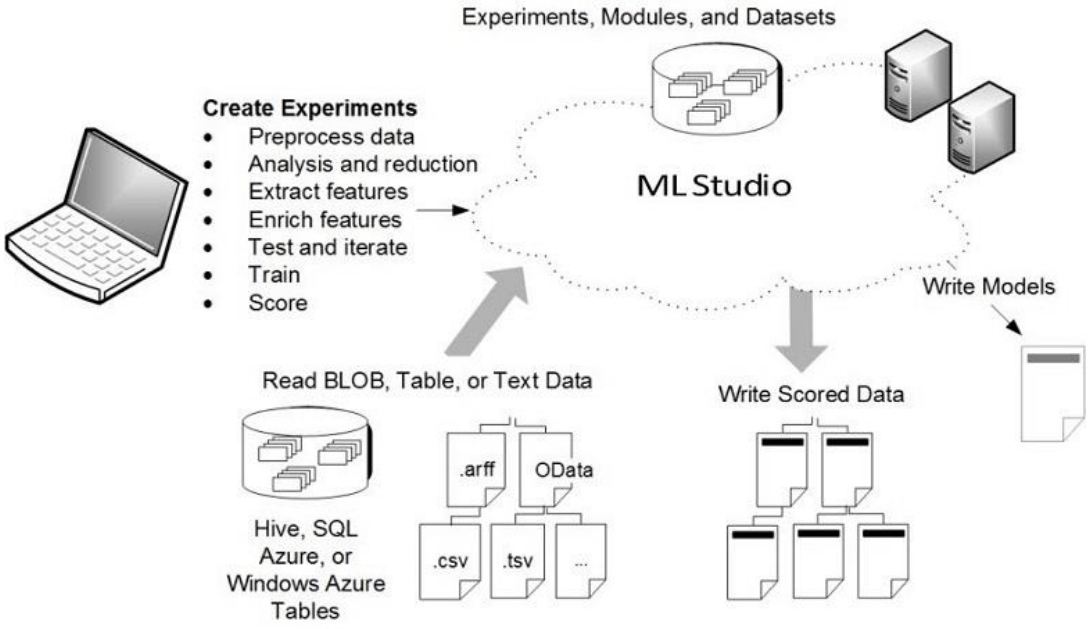
El Machine Learning o modelos de aprendizaje automático es una rama dentro de la Inteligencia Artificial que permite desarrollar sistemas que aprendan automáticamente en base a datos ya existentes, pudiendo predecir nuevos resultados, comportamientos y tendencias futuras para nuevos escenarios.

Para la construcción de modelos de aprendizaje automático se pueden utilizar diferentes softwares, en nuestro caso utilizamos Azure Machine Learning Studio de Microsoft; es un entorno de Azure que permite crear aplicaciones arrastrando y moviendo de forma muy visual. Se usa directamente desde el explorador de forma sencilla, pero conservando toda la potencia de Machine Learning (https://mvpcluster.com/microsoft-azure-machine-learning-studio/).

La versión clásica de Azure Machine Learning Studio ofrece un espacio de trabajo interactivo, fácil de construir, testear, e iterar con modelos predictivos. Se arrastran los sets de

datos en los módulos de análisis y se conectan para realizar los experimentos. Los experimentos se pueden guardar, se pueden editar y volver a correr.

Una vez que el modelo está listo, se puede publicar con el servicio web y de esta forma el experimento puede estar disponible para otros. No requiere programación, sino que se conectan los datos y módulos de forma visual para construir los modelos.



Medidas de performance de los modelos de aprendizaje automático utilizadas.

Para la evaluación de precisión de los modelos se utilizará accuracy, matriz de confusión y curva AUC-ROC.

La curva ROC nos dice qué tan bueno puede distinguir el modelo entre dos cosas, por ejemplo, si un cliente nos va a atender el teléfono o no. Mejores modelos pueden distinguir con precisión entre los dos, mientras que un modelo pobre tendrá dificultades para distinguir entre los dos.

El AUC es el área bajo la curva ROC. Este puntaje nos da una buena idea de qué tan bien funciona el modelo.

Cuando dos curvas no se superponen en absoluto, el modelo tiene una medida ideal de separación. Es perfectamente capaz de distinguir entre clase positiva y clase negativa.

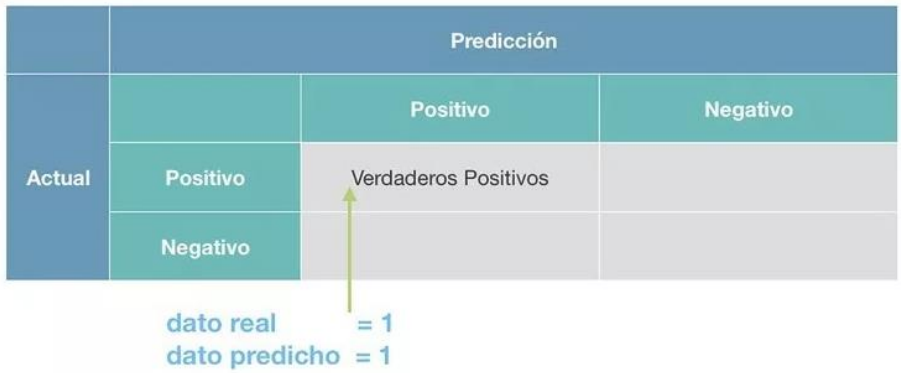
Cuando dos distribuciones se superponen, introducimos errores. Dependiendo del umbral, podemos minimizarlos o maximizarlos. Cuando AUC es 0.7, significa que hay 70% de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa.

Esta es la peor situación. Cuando el AUC es aproximadamente 0.5, el modelo no tiene capacidad de discriminación para distinguir entre clase positiva y clase negativa.

La Matriz de Confusión es una de las métricas más intuitivas y sencillas que se utiliza para encontrar la precisión y exactitud del modelo. Se utiliza para el problema de clasificación donde la salida puede ser de dos o más tipos de clases.

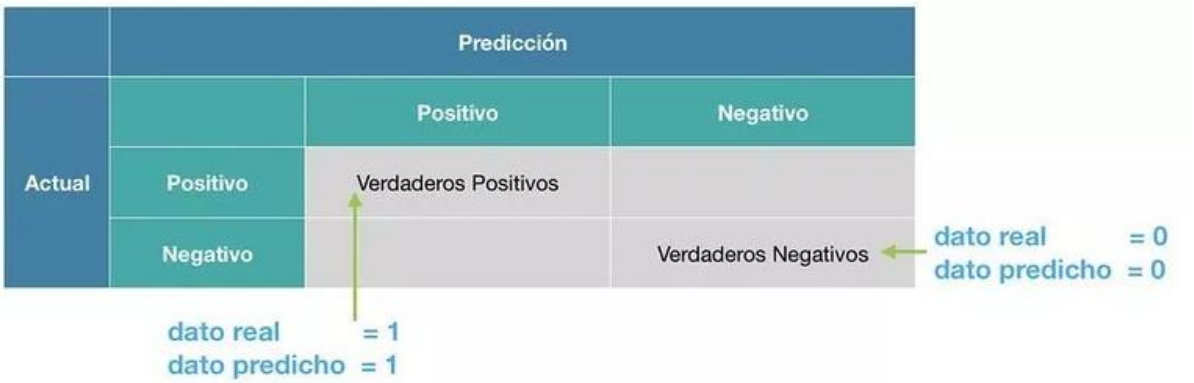
La Matriz de confusión, es una tabla con dos dimensiones, “Actual” y “Predicción”, y conjuntos de clases en ambas dimensiones. Las filas de la matriz indican la clase observada o real y las columnas indican la clase predicha.

Verdaderos Positivos (True Positives – TP)



Son los casos en los que los datos reales son 1 (Verdadero) y la predicción también es 1 (Verdadero).

Verdaderos Negativos (True Negatives – TN)



Son los casos en los que los datos reales con 0 (Falso) y el pronóstico también es 0 (Falso).

Falsos Positivos (False Positives – FP)

Son los casos en que los datos reales indica que es 0 (Falso) y la predicción indica que es 1 (Verdadero), es decir la predicción ha sido errónea. La palabra Falso es porque el modelo ha pronosticado incorrectamente y positivo porque la predicción ha sido positiva (1).

Falsos Negativos (False Negatives – FN)

Son los casos en que los datos reales indica que es 1 (Verdadero) y el pronóstico es 0 (Falso), ocasionando que la predicción ha sido incorrecta. La palabra Falso es porque el modelo ha predicho incorrectamente y negativo porque predijo que era negativa (0).

El escenario ideal que todos queremos es que el modelo dé 0 falsos positivos y 0 falsos negativos, pero ese no es el caso en la vida real, ya que cualquier modelo NO será 100% preciso en la mayoría de los casos.

La exactitud (accuracy) de la clasificación es la relación entre las predicciones correctas y el número total de predicciones. O más simplemente, con qué frecuencia es correcto el clasificador. (http://ligdigonzalez.com/)

Construcción de los modelos de aprendizaje automático. Tipos de modelos a entrenar y sus resultados.

Modelos de aprendizaje automático utilizados.

Los modelos que entrenaremos sobre ambos sets de datos para luego seleccionar aquel que nos dé el mejor resultado serán; Bayes, regresión logística, árbol de decisión, decision forest, decision jungle, redes neuronales, y SVM.

A continuación, haremos una breve descripción de cada uno.

Naive Bayes o el Ingenuo Bayes es uno de los algoritmos más simples y poderosos para la clasificación basado en el Teorema de Bayes con una suposición de independencia entre los predictores. Naive Bayes es fácil de construir y particularmente útil para conjuntos de datos muy grandes.

El clasificador Naive Bayes asume que el efecto de una característica particular en una clase es independiente de otras características. Por ejemplo, un solicitante de préstamo es deseable o no dependiendo de sus ingresos, historial de préstamos y transacciones anteriores, edad y ubicación. Incluso si estas características son interdependientes, estas características se consideran de forma independiente. Esta suposición simplifica la computación, y por eso se considera ingenua. Esta suposición se denomina independencia condicional de clase.

En caso de que se tenga una sola característica, el clasificador Naive Bayes calcula la probabilidad de un evento en los siguientes pasos:

Paso 1: calcular la probabilidad previa para las etiquetas de clase dadas.

Paso 2: determinar la probabilidad de probabilidad con cada atributo para cada clase.

Paso 3: poner estos valores en el teorema de Bayes y calcular la probabilidad posterior.

Paso 4: ver qué clase tiene una probabilidad más alta, dado que la variable de entrada pertenece a la clase de probabilidad más alta.

Las ventajas de este modelo son; es fácil y rápido predecir la clase de conjunto de datos de prueba. También funciona bien en la predicción multiclase. Cuando se mantiene la suposición de independencia, un clasificador Naive Bayes funciona mejor en comparación con otros modelos como la Regresión Logística y se necesitan menos datos de entrenamiento. Funciona bien en el caso de variables de entrada categóricas comparada con variables numéricas. (http://ligdigonzalez.com/)

La Regresión Logística es un método estadístico para predecir clases binarias. El resultado o variable objetivo es de naturaleza dicotómica. Dicotómica significa que solo hay

dos clases posibles. Por ejemplo, se puede utilizar para problemas de detección de cáncer o calcular la probabilidad de que ocurra un evento.

La Regresión Logística es uno de los algoritmos de Machine Learning más simples y más utilizados para la clasificación de dos clases. Es fácil de implementar y se puede usar como línea de base para cualquier problema de clasificación binaria. La Regresión Logística describe y estima la relación entre una variable binaria dependiente y las variables independientes.

La Regresión Logística lleva el nombre de la función utilizada en el núcleo del método, la función logística es también llamada función Sigmoide. Esta función es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1.

Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO. Por su parte si el resultado es 0.75, podemos decir en términos de probabilidad como, hay un 75% de probabilidades de que el paciente sufra cáncer.

En resumen, la Regresión Logística es el algoritmo de Machine Learning más famoso después de la Regresión Lineal, es un algoritmo simple que se puede utilizar para tareas de clasificación binarias y multivariadas. (http://ligdigonzalez.com/)

Un árbol de decisión tiene una estructura similar a un diagrama de flujo donde un nodo interno representa una característica o atributo, la rama representa una regla de decisión y cada nodo u hoja representa el resultado. El nodo superior de un árbol de decisión se conoce como nodo raíz.

La idea básica detrás de cualquier problema de árbol de decisión es la siguiente; seleccionar el mejor atributo utilizando una medida de selección de atributos o características. Hacer de ese atributo un nodo de decisión y dividir el conjunto de datos en subconjuntos más pequeños. Comenzar la construcción del árbol repitiendo este proceso recursivamente para cada atributo hasta que una de las siguientes condiciones coincida; todas las variables pertenecen al mismo valor de atributo o ya no quedan más atributos o no hay más casos.

La medida de selección de atributos es una heurística para seleccionar el criterio de división que divide los datos de la mejor manera posible. También se conoce como reglas de partición porque nos ayuda a determinar puntos de ruptura para conjunto en un nodo dado.

Esta medida proporciona un rango a cada característica, explicando el conjunto de datos dado. El atributo de mejor puntuación se seleccionará como atributo de división. En el caso de un atributo de valor continuo, también es necesario definir puntos de división por las ramas.

Las medidas de selección más populares son la ganancia de información, la relación de ganancia y el índice de Gini.

Algunos de las ventajas que tiene este algoritmo son las siguientes; los árboles de decisión son fáciles de interpretar y visualizar. Puede capturar fácilmente patrones no lineales. Requiere menos preprocesamiento de datos por parte del usuario, por ejemplo, no es necesario normalizar las columnas. Se puede utilizar para ingeniería de características, como la predicción de valores perdidos, adecuada para la selección de variables. El árbol de decisión no tiene suposiciones sobre la distribución debido a la naturaleza no paramétrica del algoritmo. (http://ligdigonzalez.com/)

Los Bosques Aleatorios es un algoritmo de aprendizaje supervisado. Puede utilizarse tanto para la clasificación como para la regresión. También es el algoritmo más flexible y fácil de usar. Un bosque está compuesto de árboles. Se dice que cuantos más árboles tenga, más robusto será el bosque. Los Bosques Aleatorios crea árboles de decisión a partir de muestras de datos seleccionados al azar, obtiene predicciones de cada árbol y selecciona la mejor solución mediante votación. También proporciona un indicador bastante bueno de la importancia de la característica.

Los Bosques Aleatorios tienen una variedad de aplicaciones, tales como motores de recomendación, clasificación de imágenes y selección de características. Se puede utilizar para clasificar a los solicitantes de préstamos, identificar actividades fraudulentas y predecir enfermedades.

Técnicamente es un método de conjunto, basado en el enfoque de dividir y conquistar, de árboles de decisión generados en un conjunto de datos dividido al azar. Los árboles de decisión individuales se generan utilizando un indicador de selección de atributos, como la ganancia de información, la relación de ganancia y el índice Gini, para cada atributo. Cada árbol depende de una muestra aleatoria independiente. En un problema de clasificación, cada árbol vota y se elige la clase más popular como resultado final. Es más simple y más potente en comparación con otros algoritmos de clasificación no lineal.

Funciona en cuatro pasos; construir un árbol de decisión para cada muestra, obtener un resultado de predicción de cada árbol de decisión, realizar una votación por cada resultado previsto y seleccionar el resultado de la predicción con más votos como predicción final.

Las ventajas de este modelo son; los Bosques Aleatorios se consideran un método muy preciso y robusto debido al número de árboles de decisión que participan en el proceso. No sufre el problema del sobreajsute. La razón principal es que toma el promedio de todas las predicciones, lo que anula los sesgos. El algoritmo puede utilizarse tanto en problemas de

clasificación como de regresión. Los Bosques Aleatorios también pueden manejar los valores que faltan. Hay dos maneras de manejarlos: usando valores medianos para reemplazar variables continuas, y calculando el promedio ponderado por proximidad de los valores faltantes. (http://ligdigonzalez.com/)

Las Máquinas Vectores de Soporte clasificación ofrece una precisión muy alta en comparación con otros clasificadores como la Regresión Logística y los Árboles de Decisión. Es conocido por su truco de kernel para manejar espacios de entrada no lineales. Se utiliza una variedad de aplicaciones tales como detección de rostros, detección de intrusos, clasificación de correos electrónicos, artículos de noticias y páginas web, entre otros.

El SVM, por sus siglas en inglés, construye un hiperplano en un espacio multidimensional para separar las diferentes clases. El SVM genera un hiperplano óptimo de forma iterativa, que se utiliza para minimizar un error. La idea central de SVM es encontrar un hiperplano marginal máximo que mejor divida el conjunto de datos en clases.

El objetivo principal es segregar el conjunto de datos de la mejor manera posible. La distancia entre los puntos más cercanos se conoce como el margen. El objetivo es seleccionar un hiperplano con el máximo margen posible entre vectores de soporte en el conjunto de datos dado.

Para buscar el mejor hiperplano se siguen los siguientes pasos: generar hiperplanos que segreguen las clases de la mejor manera. Como podemos observar en la figura tenemos tres hiperplanos que separan los datos, pero solamente uno de ellos está separando las dos clases correctamente, el resto tienen mayor error de clasificación.

Seleccionar el hiperplano correcto con la máxima segregación de los puntos de datos más cercanos.

Los clasificadores de Máquinas de Vectores de Soporte ofrecen una buena precisión y realizan predicciones más rápidas en comparación con el algoritmo de Naive Bayes. También utilizan menos memoria porque utilizan un subconjunto de puntos de entrenamiento en la fase de decisión. Este algoritmo funciona bien con un claro margen de separación y con un espacio dimensional elevado.

Las Máquinas de Vectores de Soporte no son adecuadas para grandes conjuntos de datos debido a su alto tiempo de formación y también requiere más tiempo de formación en comparación con Naive Bayes. Funciona mal con clases superpuestas y también es sensible al tipo de núcleo utilizado. (http://ligdigonzalez.com/)

Por último, se utiliza el modelo de redes neuronales para salida de clasificación de dos clases. Una red neuronal es un conjunto de capas interconectadas. Las entradas son la primera

capa y se conectan a una capa de salida mediante un grafo acíclico que consta de nodos y aristas ponderadas.

Entre las capas de entrada y salida puede insertar varias capas ocultas. La mayoría de las tareas predictivas pueden realizarse fácilmente con solo una o varias capas ocultas. Sin embargo, las investigaciones recientes han demostrado que las redes neuronales profundas (DNN) con muchas capas pueden ser eficaces en tareas complejas, como en el reconocimiento de imágenes o de voz. Las capas sucesivas se usan para modelar los crecientes niveles de profundidad semántica.

La relación entre entradas y salidas se aprende mediante el entrenamiento de la red neuronal con los datos de entrada. La dirección del grafo procede desde las entradas a través de la capa oculta y hacia la capa de salida. Todos los nodos de una capa se conectan mediante las aristas ponderadas a los nodos de la capa siguiente.

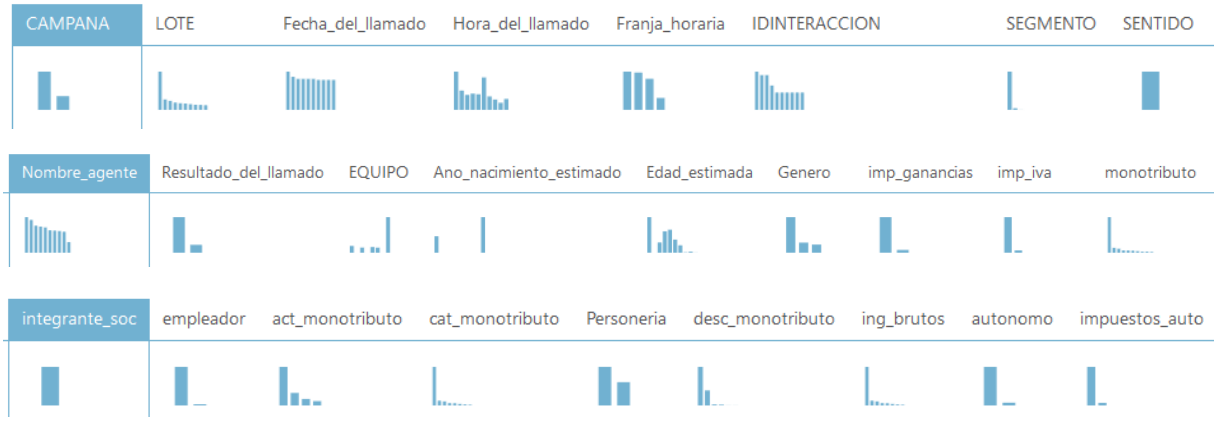
Para calcular la salida de la red para una entrada determinada, se calcula un valor en cada nodo en las capas ocultas y en la capa de salida. El valor se establece calculando la suma ponderada de los valores de los nodos de la capa anterior. A continuación, se aplica una función de activación a esa suma ponderada. (https://docs.microsoft.com/es-es/azure/machine-learning/algorithm-module-reference/two-class-neural-network)

Implementación del primer modelo de aprendizaje automático y sus resultados

Como se mencionó anteriormente se parte de una tabla que contiene los resultados de los llamados realizados por un call center entre Enero y Mayo del 2019.

La primera predicción por realizar consiste en determinar la probabilidad de que la persona atienda o no atienda el llamado. Para ello se seleccionan sólo aquellas columnas que contienen información previa a la realización del llamado. Se obtiene una tabla con 26 columnas y 53.746 filas.

En los gráficos a continuación se observan los nombres de las columnas utilizadas:



Un problema que surge con la base de datos es que no está balanceada, las probabilidades de ocurrencia tanto de que nos atiendan el llamado, cómo de éxito de la venta son muy bajas.

En estos casos, un modelo estimado sobre la base de datos completa tiene menos oportunidad de reconocer diferencias que sobre una base de datos balanceada.

Para conseguir nuestro objetivo podemos hacer uso del sobremuestreo (oversampling), submuestreo (undersampling) o ponderación (weighting).

En este caso, el software utilizado nos ofrece la funcionalidad llamada “SMOTE” (*Synthetic Minority Oversampling Technique*). Esta es una técnica estadística que nos permite incrementar la cantidad de casos en el set de datos de forma de balancearlo. El módulo genera nuevas instancias de la clase minoritaria. Este módulo no genera ni cambia el numero de la clase mayoritaria.

Una vez balanceada la base se divide la misma en 80% para entrenamiento de los modelos y 20% para las pruebas de los mismos.

El software nos permite aplicar un operador llamado “Tune Model Hyperparameters”. El objetivo es determinar los hiperparámetros óptimos para un modelo de Machine Learning. El módulo compila y prueba varios modelos con diferentes combinaciones de configuraciones. Compara las métricas de todos los modelos para obtener las combinaciones de valores.

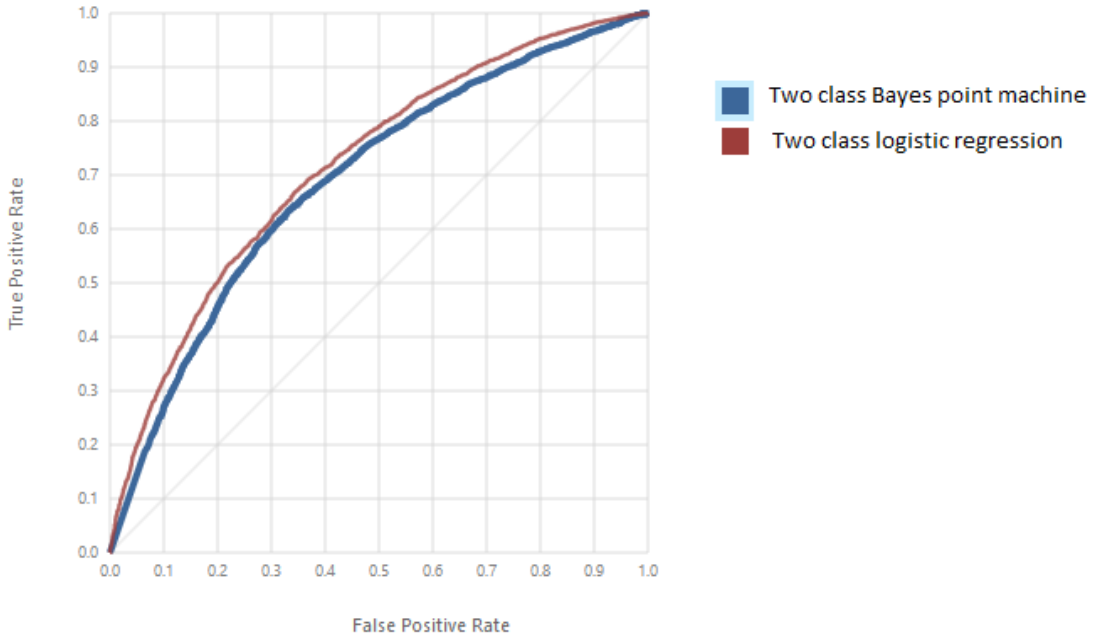
Básicamente, este módulo realiza un barrido de parámetros en la configuración de parámetros especificada. Aprende un conjunto óptimo de hiperparámetros, que puede ser diferente para cada árbol de decisión, conjunto de datos o método de regresión específico. El proceso de búsqueda de la configuración óptima a veces se denomina de optimización.

El módulo admite el siguiente método para encontrar la configuración óptima de un modelo: entrenamiento y optimización integrados. En este método, se configura el conjunto de parámetros que se van a utilizar. A continuación, se deja que el módulo recorra en iteración varias combinaciones. El módulo mide la precisión hasta que encuentra un modelo "mejor". (https://docs.microsoft.com/es-es/azure/machine-learning/algorithm-module-reference/tune-model-hyperparameters)

A continuación, se seleccionan los modelos que deseamos entrenar. El software nos permite ejecutar de a dos modelos por vez comparando los resultados de estos en una sola ejecución.

Se seleccionan distintos modelos y se van comparando los resultados de uno y otro con las medidas de exactitud, curva ROC y AUC tal como se describió en el apartado anterior.

Resultados de los modelos aplicados a los datos:



Bayes:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
6580	2244	0.681	0.784	0.5	0.687

False Positive	True Negative	Recall	F1 Score
1812	2065	0.746	0.764

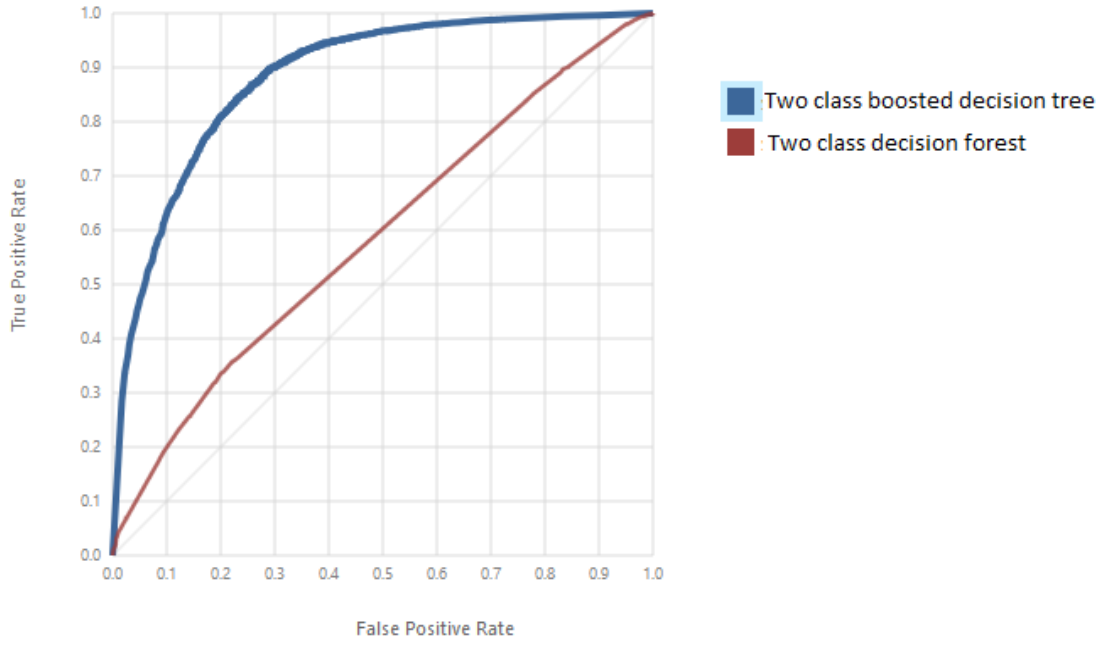
Positive Label	Negative Label
No atiende	Atiende

Regresión logística:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
7961	863	0.723	0.750	0.5	0.718

False Positive	True Negative	Recall	F1 Score
2654	1223	0.902	0.819

Positive Label	Negative Label
No atiende	Atiende



Árbol de decisión:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
7903	921	0.839	0.876	0.5	0.884

False Positive	True Negative	Recall	F1 Score
1122	2755	0.896	0.886

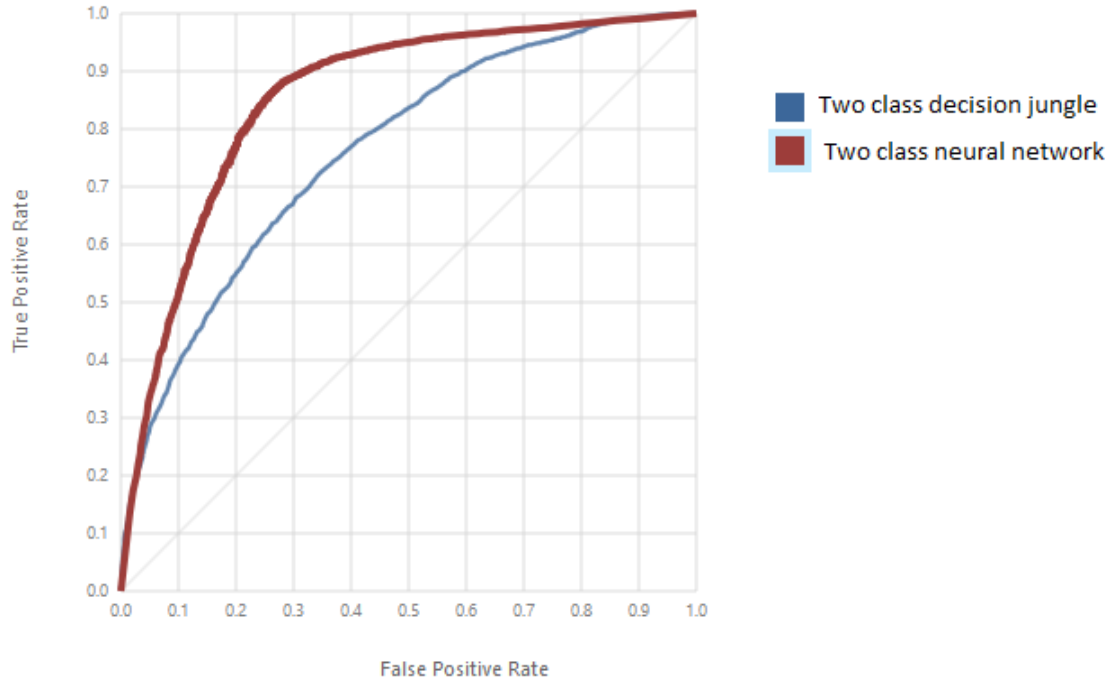
Positive Label	Negative Label
No atiende	Atiende

Decision forest:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
8781	43	0.696	0.697	0.5	0.588

False Positive	True Negative	Recall	F1 Score
3819	58	0.995	0.820

Positive Label	Negative Label
No atiende	Atiende



Decision jungle:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
8499	325	0.736	0.737	0.5	0.761

False Positive	True Negative	Recall	F1 Score
3032	845	0.963	0.835

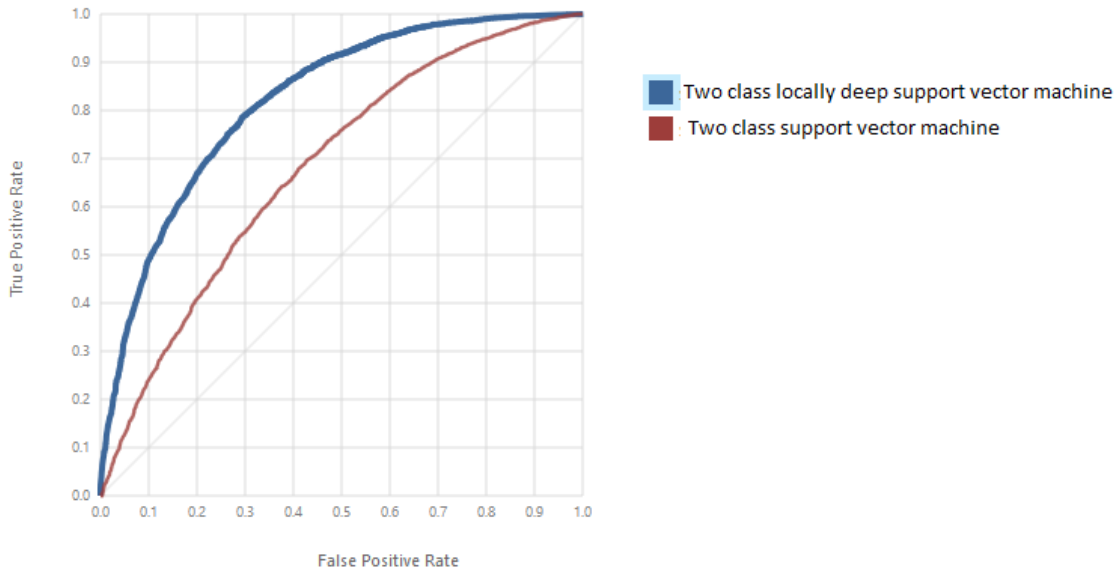
Positive Label	Negative Label
No atiende	Atiende

Red Neuronal:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
6864	1960	0.784	0.897	0.5	0.854

False Positive	True Negative	Recall	F1 Score
787	3090	0.778	0.833

Positive Label	Negative Label
No atiende	Atiende



Locally Deep SVM:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
7936	888	0.791	0.818	0.5	0.822

False Positive	True Negative	Recall	F1 Score
1771	2106	0.899	0.857

Positive Label	Negative Label
No atiende	Atiende

SVM:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
8232	592	0.723	0.737	0.5	0.682

False Positive	True Negative	Recall	F1 Score
2932	945	0.933	0.824

Positive Label	Negative Label
No atiende	Atiende

Implementación del segundo modelo de aprendizaje automático y sus resultados.

En base a los resultados obtenidos en el primer modelo de predicción de la probabilidad de respuesta/no respuesta de los llamados a realizar por el call center podemos observar que el modelo que arroja los mejores resultados es el de árbol de decisión.

Para determinar cuáles de las variables ingresadas como entrada al modelo son relevantes para la predicción se utiliza un operador dentro del software llamado “Permutation Feature Importance”

El mismo se utiliza para calcular un conjunto de puntuaciones de importancia de las características del conjunto de datos. Estas puntuaciones se usan para ayudar a determinar las mejores características que se deben usar en un modelo.

En este módulo, los valores de las características se ordenan aleatoriamente, una columna cada vez. El rendimiento del modelo se mide antes y después.

Las puntuaciones que devuelve el módulo representan el cambio en el rendimiento de un modelo formado, después de la permutación. Las características importantes suelen ser más sensibles al proceso de orden aleatorio y, por tanto, se obtendrán mayores puntuaciones de importancia. El módulo captura el grado de influencia que tiene cada característica en las predicciones del modelo.

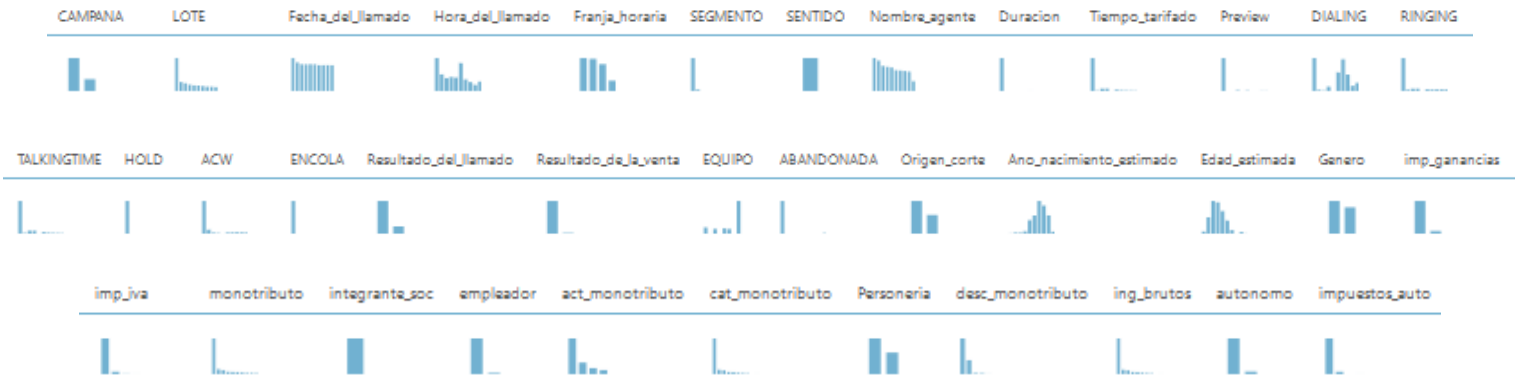
En este primer modelo, de predicción de atención del llamado tenemos el siguiente resultado:

Feature	Score
Fecha_del_llamado	0.093142
LOTE	0.080545
EQUIPO	0.079521
monotributo	0.059208
cat_monotributo	0.0485
Hora_del_llamado	0.04787
Franja_horaria	0.041099
Nombre_agente	0.027321
Ano_nacimiento_estimado	0.027242
Edad_estimada	0.025116
desc_monotributo	0.00685
CAMPANA	0.005748
SEGMENTO	0.005354
Genero	0.004724
act_monotributo	0.004173
Personeria	0.001181
imp_ganancias	0.000709
empleador	0.00063
imp_iva	0.000394
SENTIDO	0
integrante_soc	0
ing_brutos	0
autonomo	-0.000079
impuestos_auto	-0.000079

Cómo se puede observar, las variables sentido, integrante de sociedad, ingresos brutos, autónomo e impuestos autónomo no son relevantes en este modelo. Las demás sí lo son y ayudan a determinar la probabilidad de que la persona nos atienda o no el llamado que estamos realizando.

La segunda predicción por realizar consiste en determinar la probabilidad de que la persona acepte o no la oferta de crédito que está recibiendo. Para ello se seleccionan todas las columnas de la tabla, que nos dan información además de las características del llamado que recibió.

En los gráficos a continuación se observan los nombres de las columnas utilizadas:

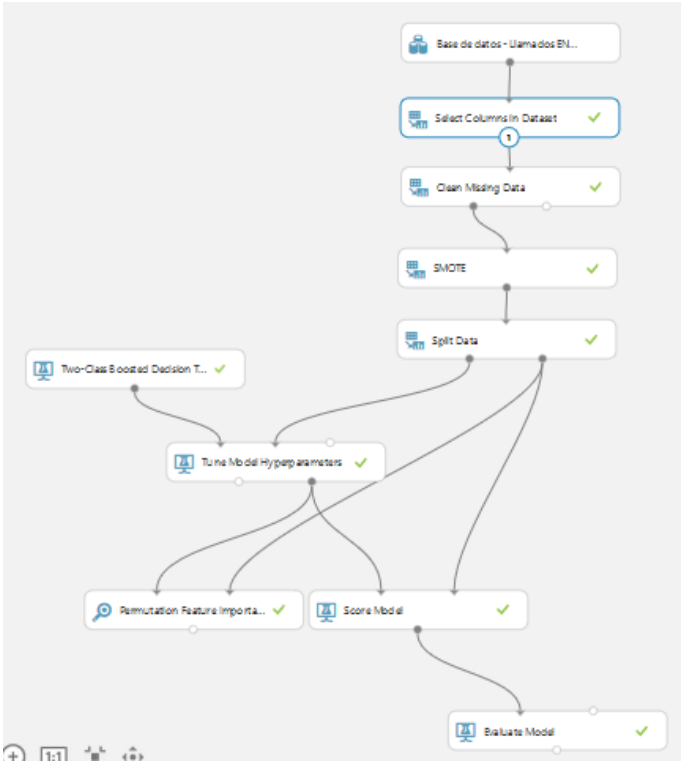


Este modelo es de utilidad sólo para medir qué variables determinan el éxito o no en la venta del crédito, ya que utilizamos como variables datos que sólo obtendremos luego de haber hablado con el potencial cliente.

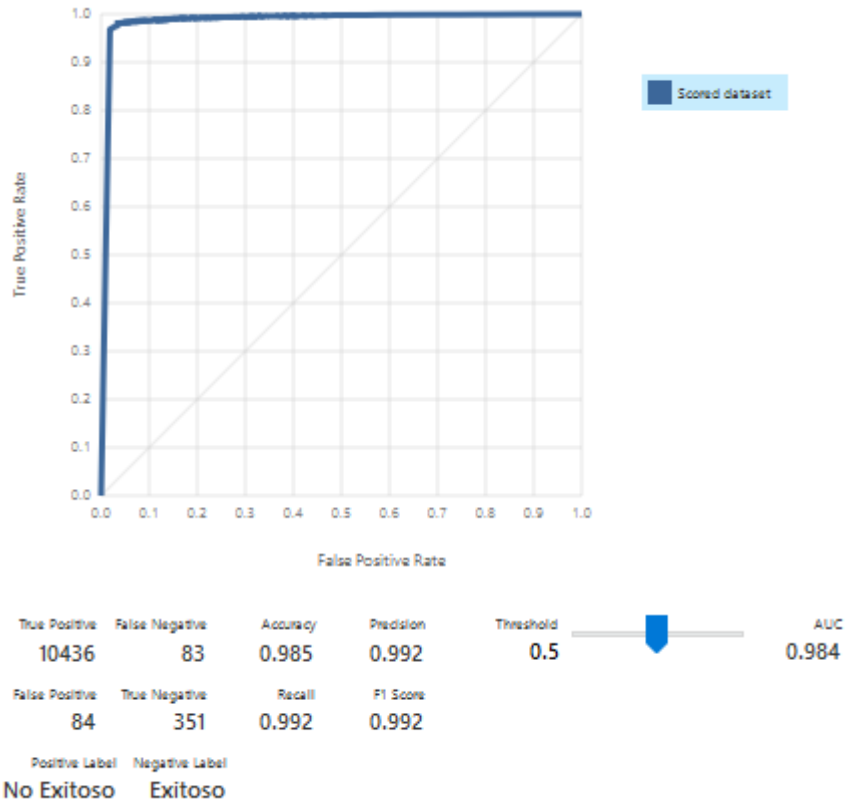
Es decir, que el primer modelo predictivo descrito con anterioridad es el que nos servirá a priori para ordenar y seleccionar los llamados a realizar por el call center y de esta forma ayudar a que realicen un trabajo más eficiente.

Este segundo modelo nos sirve para evaluar los resultados. Para ello no probaremos con diferentes modelos como hicimos en el caso anterior, si no que utilizaremos un modelo de árbol que nos permite luego analizar las variables que pesan en el mismo y no se comporta como "caja negra" como si sucede con otros como redes neuronales.

Se utiliza en este caso, también, Microsoft Azure Machine Learning Studio y se programa la secuencia a continuación:



El modelo se mide de la misma forma que el caso anterior obteniendo los siguientes resultados:



En el análisis de las variables que aportan al modelo, la principal es claramente que la persona atienda o no el llamado, lo que confirma que el modelo principal a utilizar es el anterior, si se logra que el call center llame en su mayoría a las personas que van a atender el llamado el principal factor de éxito estará logrado.

Feature	Score
Resultado_del_llamado	0.019901
HOLD	0.012233
TALKINGTIME	0.008216
Duracion	0.00776
Fecha_del_llamado	0.004199
Tiempo_tarifado	0.004017
ACW	0.003834
Nombre_agente	0.00283
Ano_nacimiento_estimado	0.000822
LOTE	0.000639
RINGING	0.000365
monotributo	0.000365
Origen_corte	0.000274
Franja_horaria	0.000183
DIALING	0.000183
Genero	0.000183
CAMPANA	0.000091
Preview	0.000091
empleador	0.000091
SEGMENTO	0
SENTIDO	0
ENCOLA	0
ABANDONADA	0
imp_ganancias	0
imp_iva	0
integrante_soc	0
cat_monotributo	0
Personeria	0
ing_brutos	0
autonomo	0
impuestos_auto	0
act_monotributo	-0.000091
desc_monotributo	-0.000091
EQUIPO	-0.000183
Hora_del_llamado	-0.000365
Edad_estimada	-0.000548

Conclusión

Decidir con datos permitió a la empresa crecer de forma exponencial en los últimos años, en los primeros 7 años la empresa alcanzó treinta mil clientes vigentes y se había estancado en ese número, y en los últimos 3 años, a partir del manejo de los datos y contar con más y mejor información se logró medir el crecimiento, incorporar objetivos específicos asociados a esto en el modelo de compensaciones de los ejecutivos y crecer diez mil clientes entre Junio de 2017 y Diciembre 2019 (cierre de año proyectado).

A su vez, los modelos predictivos de propensión de default y de propensión de fuga desarrollados basados en los datos internos de la empresa permitieron precalificar clientes a los que dejó de ser necesario evaluarlos en el lugar de trabajo por parte de la fuerza comercial, esto trae consigo una importante baja en los costos operativos, y una mayor capacidad de colocación de créditos por mes.

De lo anterior se concluye que los datos bien almacenados y administrados generan información valiosa que puede llevar a las empresas a liderar en su sector, a tener procesos más eficientes, colaboradores más satisfechos ya que tienen claridad respecto de lo que se espera de ellos y de lo que están haciendo bien y cuáles son sus brechas de mejora y bajar costos.

Con el dataset trabajado se logró confeccionar dos modelos predictivos que pueden mejorar la gestión de ventas de la organización a través de su call center.

En base a lo aquí trabajado se recomienda utilizar los modelos sobre nuevas bases a gestionar ordenando y clasificando los llamados a realizar. En base a los nuevos resultados se pueden ajustar y retroalimentar los modelos.

A su vez, se podrían incorporar nuevas fuentes de información a las bases como por ejemplo datos provenientes de redes sociales. En este y en todos los casos se debe cuidar la privacidad de los datos, la ética en la utilización de estos, y la autorización para acceder a los mismos.

Referencias bibliográficas

1- Microfinanzas bajo la lupa – Revista Forbes – Tornero Lucia – 2018 - http://www.forbesargentina.com/microfinanzas-bajo-la-lupa/

2- Un mundo sin pobreza – Yunus, Muhammad – Editorial Paidós – 2008

3- La inclusión financiera es un factor clave para reducir la pobreza e impulsar la prosperidad. – Abril 2018 - https://www.bancomundial.org/es/topic/financialinclusion/overview

4- Sitio web Provincia Microcréditos S.A. - https://www.provinciamicrocreditos.com/quienes-somos/

5- Grameen Bank What is Microcredit. https://www.grameen-info.org/what-is-microcredit/

6- Programa de las Naciones Unidas para el Desarrollo Microfinanzas en la Argentina – 1ra edición – Buenos Aires. PNUD, 2005. https://avanzar.org.ar/wp-content/uploads/2018/02/Microfinanzas-en-Argentina.pdf

7- Fundación Avanzar - https://avanzar.org.ar/

8- ALGORITMOS DE APRENDIZAJE: KNN&KMEANS [Inteligencia en Redes de Telecomunicación] Cristina García Cambronero e Irene Gómez Universidad Carlos III de Madrid. http://www.it.uc3m.es/~jvillena/irc/practicas/08-09/06.pdf

9- Sitio web Microsoft https://products.office.com/es-ar/project/project-management-software

10- SQL Server 2014. Administración de una base de datos transaccional con SQL Management Studio. Jerome Gabillaud. Ediciones ENI. Julio 2015.

11- https://mvpcluster.com/microsoft-azure-machine-learning-studio/

12- Sitio web de Microsoft https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio

13- http://ligdigonzalez.com/

Anexos

Anexo 1.
Descripción de las variables de la tabla completa.

Nombre de la columna	Descripción	Comentario
ID	Se agrega valor numérico de ID para identificar de forma única cada registro	
CAMPAÑA	Tipo de campaña a la que pertenece el registro.	
LOTE	Clasificación del lote de la campaña	
Fecha del llamado	Fecha y hora en la que se realizó el llamado	
Hora del llamado	Se extrae de la columna fecha del llamado la hora	De creación manual
Franja horaria	Se dividen en franjas los horarios de los llamados: 8-10, 11-13, 14-17, 18-20	De creación manual
IDINTERACCION	ID interacción, posee valores duplicados ya que identifica cliente y acción y no cada intento de forma única	
SEGMENTO	Valor numérico de 1 a 20	
SENTIDO	Indica si el llamado es entrante o saliente	
NOMBRE AGENTE	Nombre del operador que hace el llamado	
DURACIÓN	duración del llamado	
TIEMPO TARIFADO	tiempo que se factura del llamado	
PREVIEW	tiempo que el operador usa para entender el caso	
DIALING	tiempo que se demora marcando	
RINGING	tiempo esperando a ser atendido	
TALKINGTIME	tiempo de conversación del llamado	
HOLD	tiempo en espera	
ACW	tiempos dentro del llamado	
ENCOLA	tiempos dentro del llamado	

TIPIFICACIÓN	clasificación del resultado del llamado	Se excluye en ambos modelos
Resultado del llamado	clasificación del resultado en "atiende" "no atiende" - primer variable a predecir	columna creada. En el segundo modelo se excluye
TIPOS TIPIFICACIÓN	tipificación del resultado del llamado - segunda variable a predecir	En el primer modelo se excluye
Mail	mail del contacto	
EQUIPO	equipo al que pertenece el operador	
ABANDONADA	0 si no se abandona el llamado 1 si el llamado resulta abandonado	
ORIGEN CORTE	si el llamado es interrumpido por el operador o por el contacto	
SITUACION	situación de CENDEU - central deudores de BCRA - donde indica la última peor situación de deuda del contacto respecto al sistema financiero - siendo 1 al día - NULL sin información - 2 o mayor significa que la persona tiene atrasos con sus deudas (información obtenida de BCRA)	columna agregada de otra fuente
Sexo	Se infiere el sexo en base al cuil en aquellos casos que comienza con 20(hombre) y 27 (mujer)	Columna creada
año de nacimiento estimado	Se estima el año de nacimiento según dni del contacto	columna creada
edad estimada	se estima edad según año de nacimiento estimado	columna creada
imp_ganancias	si el contacto está o no inscripto en impuesto a las ganancias (información obtenida de a página web de AFIP)	columna agregada de otra fuente
imp_iva	si el contacto está o no inscripto en iva (información obtenida de a página web de AFIP)	columna agregada de otra fuente
monotributo	si el contacto está o no inscripto y su categoría de monotributo (información obtenida de a página web de AFIP)	columna agregada de otra fuente

integrante_soc	si el contacto es pate de una sociedad (información obtenida de a página web de AFIP)	columna agregada de otra fuente
empleador	si el contacto tiene registrados empleados a cargo(información obtenida de a página web de AFIP)	columna agregada de otra fuente
act_monotributo	si el contacto está o no inscripto y su categoría de monotributo (información obtenida de a página web de AFIP)	columna agregada de otra fuente
cat_monotributo	si el contacto está o no inscripto y su categoría de monotributo (información obtenida de a página web de AFIP)	columna agregada de otra fuente
pers	si el contacto está inscripto como persona física (información obtenida de a página web de AFIP)	columna agregada de otra fuente
desc_monotributo	tipo de actividad en monotributo (información obtenida de a página web de AFIP)	columna agregada de otra fuente
ing_brutos	categoría de ingresos brutos en el caso de estar inscripto (información obtenida de a página web de AFIP)	columna agregada de otra fuente
autonomo	si el contacto está o no inscripto como autónomo (información obtenida de a página web de AFIP)	columna agregada de otra fuente
impuestos_auto	si está como autónomo en qué categoría (información obtenida de a página web de AFIP)	columna agregada de otra fuente

Solicitud de aprobación de PROYECTO DE TRABAJO FINAL DE ESPECIALIZACIÓN	Código de la Especialización
María Clara Accame	DNI 33422723
2019	

Título del Trabajo Final (preliminar): Ventas telefónicas de **microcréditos**. Diseño de un **modelo de predicción de la probabilidad de éxito** de venta.

Conformidad del profesional propuesto como Tutor de Trabajo Final

He revisado el proyecto y acepto la postulación como Tutor comprometiéndome a dirigir las tareas del alumno orientadas a elaborar su Trabajo Final de Especialización.

Firma del Tutor de Trabajo Final ……

Aclaración……

Datos de contacto del postulante a Tutor	
Correo electrónico	Teléfonos

Se adjunta a este formulario:

- Proyecto de Trabajo Final de Especialización
- CV del postulante a Tutor de Trabajo Final (si no fuera docente de la Especialización)

Fecha	Firma del alumno

Para uso exclusivo de la Dirección de la Especialización

Se solicita a la EEP elevar al Consejo Directivo de la FCE el pedido de aprobación de tema de Trabajo Final y designación de Tutor/a propuesto/a.

FIRMA AUTORIDAD ACADÉMICA ACLARACIÓN

FECHA

Form. PTFE v0

PRESENTAR EN LA RECEPCIÓN DE LA ESCUELA DE ESTUDIOS DE POSGRADO