


UBA FCE
Universidad de Buenos Aires
Facultad de Ciencias Económicas

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 INTEGRADOR

Desarrollo de modelos de aprendizaje automático a través de metodologías agiles.

AUTOR: LIC. SEBASTIÁN ANDRADA

MENTOR: ING. YAMILA ZAKHEM

DICIEMBRE DE 2020

Resumen

Big data es un término que representa la evolución en la capacidad de análisis de datos en términos de volumen, velocidad y variedad. Este fenómeno fue posible gracias a desarrollos computacionales que dieron soporte, pudiendo ampliar el recorrido del dato desde una analítica tradicional, limitada a la descripción y el diagnostico, hacia la analítica avanzada con la capacidad de poder predecir y prescribir escenarios.

En el presente trabajo se plantea la forma de explotar los beneficios disponibles gracias a la gestión de datos en contextos organizacionales a partir de la utilización de técnicas de aprendizaje automático. El objetivo está asociado a la optimización de la base de datos de clientes del sector bancario, dentro de un contexto en el cual las empresas del sector se están reinventando para poder adaptarse a nuevas plataformas y necesidades de usuarios.

En el primer apartado se explica problemática asociada a la gestión de datos en una organización como el Banco Galicia que actualmente se encuentra en un proceso de transformación digital. Luego se explica la metodología utilizada para la obtención de datos y las particularidades que fueron surgiendo durante el desarrollo del trabajo, y, por último, el tercer apartado desarrolla la implementación de un modelo de aprendizaje automático que persigue como resultado un marco de recomendación que sirva a las organizaciones para hacer un uso más eficiente de sus recursos aumentando su porcentaje de conversión.

Palabras clave

Gestión de datos, optimización de campañas, agile, modelos de aprendizaje automático.

Índice

Introducción

El 2 de diciembre del 2020 se realizó el lanzamiento oficial de MODO la nueva billetera virtual compatible con las aplicaciones de quince bancos, número que se espera aumente en los próximos meses. El proyecto fue encabezado por el Banco Galicia y otros dos de los principales bancos del país con la idea de formar una alianza estratégica que permita competir con el desembarco en los últimos años de las Fintech. A través de esta aplicación se puede realizar compras con código QR, generar órdenes de pago y transferir a contactos del celular sin necesidad de otros datos.

MODO es un ejemplo de cómo el sector financiero está comenzando a aprovechar el potencial que tiene la gestión de datos en contextos organizacionales, lo cual puede ser aplicado prácticamente todos sus procesos. Grandes volúmenes de datos se generan a diario en las organizaciones y el desafío con el que muchas se encuentran es encontrar la utilidad su rentabilidad. Ante este escenario las empresas están atravesando una transformación digital, reestructurando sus procesos y entendiendo a los datos como un activo estratégico para el negocio, sobre el cual se basan sus decisiones y modelos de negocio. El caso del sector financiero resulta pertinente para este trabajo porque demuestra una industria que dispone de una inmensa variedad y volumen de datos que circulan en tiempo real y la diversidad de fines para los que pueden ser utilizados no tienen límites. A partir de lo mencionado anteriormente y la necesidad de desarrollar ventajas competitivas frente a las Fintech se sustenta el presente trabajo, que tiene el objetivo de buscar una forma de explotar el potencial disponible en el universo de datos con el que cuentan los bancos.

En el primer apartado se explica la problemática asociada a la gestión de datos en una organización como el Banco Galicia que actualmente se encuentra en un proceso de transformación digital. Enuncia los principales desafíos que enfrenta y el cambio cultural que está llevando adelante. Hace una breve descripción de las características de la empresa y como esta llevando adelante su transformación hacia una compañía Data Driven explicado por uno de sus principales promotores. El segundo apartado explica la metodología utilizada para la obtención de los datos y la bibliografía necesaria para la

elaboración del trabajo, y las particularidades que fueron surgiendo durante el desarrollo del trabajo. El último apartado parte del análisis de la base de datos, el cual posee variables relacionadas con el contacto realizado al cliente y sus datos demográficos. A partir del análisis de las variables se busca obtener información que sea de utilidad para el desarrollo de futuras campañas y que permita hacer un uso más eficiente de los recursos. El objetivo puntual es brindar herramientas para mejorar el porcentaje de conversión de campañas y para lograr eso se pretende desarrollar un modelo que establezca las mejores condiciones de contactos para perfiles establecidos de clientes. La selección se realiza en base al algoritmo regresión logística que permite obtener las variables que mayor influencia tienen en un resultado.

El análisis y la exploración de datos se lleva adelante con los servicios de Microsoft Azure Machine Learning Studio y el programa de visualización Power Bi, que permiten tener una visión general del dataset y comenzar a analizar las relaciones entre las variables. También se puede verificar la frecuencia y conocer el porcentaje de conversión del dataset, el cual resulta de vital importancia para el proyecto ya que es lo que se desea optimizar. Para los más de 500 mil contactos la empresa logro 9.939 ventas indicando una performance menor al 2%. Con ese resultado como disparador se plantea la necesidad de desarrollar modelos de aprendizaje automático que lleven a una gestión eficiente de los datos con los cuales cuenta una entidad bancaria y de esta manera demostrar cómo dichos modelos pueden transformar la gestión de las organizaciones. Con los resultados obtenidos se busca cerrar el trabajo aportando una visión de cómo la minería de datos está transformando la visión estratégica de las organizaciones, siendo fundamental para el desarrollo de los objetivos su correcta implementación.

1. Gestión de datos en contextos organizacionales

1.1 Descripción de la organización.

Banco Galicia

La visión del Banco Galicia está definida por su directorio y marca el camino para el accionar de sus colaboradores, “Ser el mejor Banco Universal de la Argentina: el

preferido por los clientes y los colaboradores, y el que genera más rentabilidad[1].” La misma deja en claro la intención de convertirse en un Banco al alcance de todos a través de distintas plataformas, también buscando captar los mejores talentos del mercado y siendo rentable para sus inversores. Sus acciones cotizan en la bolsa de Wall Street y siendo un Banco que opera exclusivamente en el país está posicionado como uno de los principales de la industria financiera, lo que marca un especial interés y dependencia en las políticas a nivel local. En el último año y medio, a partir de la situación económica desfavorable, sus acciones medidas en dólares perdieron más de 4 veces su valor, similar resultado que las demás compañías nacionales del sector.

Sumado a lo dicho en el párrafo anterior, la llegada del fenómeno Covid-19 implico grandes desafíos para la dirección de la organización, la cual tuvo que adaptarse a un nuevo marco macroeconómico con fuertes regulaciones y medidas que afectan el desempeño su negocio. Una de las ventajas que tuvo al momento de encarar estos desafíos es la transformación digital que está atravesando el banco desde el año 2017. La incorporación de tecnologías emergentes que se venía llevando a cabo fueron de gran utilidad para la reestructuración del negocio. Adicionalmente se comenzó a trabajar dentro de una metodología Agile, la cual brinda herramientas que contribuyen a la adaptación de grandes estructuras ante contextos de gran incertidumbre.

La incorporación de nuevas tecnologías orientadas a la gestión de datos no fue sencilla, al igual que muchas empresas los bancos se encontraron con la dificultad de tener que adaptar sus recursos tecnológicos y humanos para poder manejar el gran volumen de datos del que disponían. Provost and Fawcett (2013) advierten “The volume and variety of data have far outstripped the capacity of manual analysis, and in some cases have exceeded the capacity of conventional databases” (p. 51).

La agilidad en su estructura organizacional junto con la incorporación de la gestión de datos como activo estratégico dentro del marco de transformación digital fundamentan el motivo de elección de esta organización para el presente trabajo. Puntualmente se va a

[1]http://www.galiciasustentable.com.ar/banca/online/sustentable/web/BancoGaliciaSustentable/nuestragestionsustentable/Vision-Mision-Valores

hacer foco en la obtención de los datos y la explotación de estos para optimizar el desarrollo de campañas de venta.

1.2 Gestión de datos por parte de la organización.

En contextos de gran incertidumbre como los que se están atravesando actualmente las decisiones de negocio representan un riesgo importante, ya que al no tener un panorama claro acerca de cuál va a ser la realidad económica de un país es difícil poder predecir si las acciones van a ser correctas o no. Dentro de este marco los directivos de las organizaciones necesitan aplicar todas las herramientas a su alcance para poder reducir ese riesgo. Una encuesta[2] realizada por el MIT en colaboración con IBM en la que participaron más de 3000 ejecutivos de empresas de 100 países distintos demostraba que las empresas que tomaban decisiones basadas en datos tenían un rendimiento cinco veces mayor que las que no. Es allí donde reside la importancia de una correcta gestión de datos, no como un fin en sí mismo, sino que es una herramienta que sirve de guía y respaldo a los directivos al momento de tomar decisiones. Provost y Fawcett (2013) definen al proceso de data driven como la práctica de basar las decisiones en análisis de los datos y no solo en la pura intuición.

Esto plantea un nuevo desafío para rol de los directivos que, lejos de perder protagonismo, cuentan con la responsabilidad de aprovechar el poder del Big data para trazar la visión de la organización y guiar a la misma a explotar su máximo potencial. Su rol es trasladar la importancia de los datos a los empleados y stakeholders para obtener una visión compartida. Andrew McAfee and Erik Brynjolfsson (2012) argumentan en su artículo para la revista de Harvard lo siguiente:

> Because of big data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decisión making and performance. […] We can make better predictions and smarter decisions.

[2] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, N. Kruschwitz (2011), Mit Sloan Management Review, USA.

Avanzando en el proceso de transformación digital el rol de los datos toma un rol protagónico hacia la búsqueda de convertirse en una organización Data Driven, Fernando Raverta (2020), Chief Data Officer de Banco Galicia[3] lo explica de la siguiente manera:

> Ser Data Driven es un habilitador clave para capturar el valor de nuevas iniciativas de negocio, como por ejemplo en la propuesta de valor a nuestros clientes, generando productos simples y personalizados, mejorando el modelo de atención […].

A su vez una estructura ágil es fundamental para poder aprovechar el valor de los datos en entornos dinámicos y cambiantes como el actual. El avance de la tecnología ha permitido el surgimiento de nuevos competidores sustitutos para un sector que no conocía esos desafíos. Para competir con las Fintech los bancos deben aprovechar el volumen de datos que tienen de sus clientes para poder satisfacer la demanda de consumidores cada vez más adaptados a operar de manera digital. Con respecto a esto Raverta (2020) observa:

> Sólo considerando los datos como un activo estratégico, los bancos podremos desarrollar, adaptar o crear nuevos productos y servicios situando al cliente como el epicentro. Será el comportamiento de los clientes los que condicionen cómo tendrá que ser esta adaptación.

Como se mencionó párrafos atrás para el desarrollo de este trabajo el eje va a estar sobre datos demográficos y de contacto sobre clientes de la base de un banco. La implementación se realizará sobre una base estática, pero en una aplicación real dentro de una organización, la posibilidad de obtener retroalimentación sobre los resultados de los modelos aplicados se traduce en una mejora de estos.

1.3 Problemática de la organización y la gestión de los datos.

Provost y Fawcett (2013) plantean un paralelismo en la adopción del Big Data por parte de las organizaciones en distintas fases, así como fue con la adopción de los servicios Web. En una primera etapa la adopción del Big Data implica tener la capacidad de procesar grandes volúmenes de datos y establecer una estructura que de soporte al

[3] https://www.tynmagazine.com/banco-galicia-camino-a-ser-una-organizacion-data-driven/

negocio. La segunda etapa que denominan Big Data 2.0 llega una vez que la organización es capaz de procesar esos grandes volúmenes y se comienza a plantear que es lo que puede hacer para explotarlos y como mejorar sus procesos a través de ellos. En el Banco Galicia se dio una situación similar al momento en que se comenzaron a identificar y almacenar las distintas fuentes de datos. Al ver su dimensión el siguiente paso fue analizar cómo se podían aprovechar para mejorar el negocio y utilizarlo como una ventaja competitiva. En la encuesta realizada por el MIT que se hizo referencia anteriormente un 40% de los entrevistados indico que el principal obstáculo para la adopción de analytics para mejorar el negocio es la falta de entendimiento de cómo usar la información que se obtiene de los datos.

Entonces el cambio hacía una organización Data Driven plantea desafíos para todos los integrantes de la organización, así como también un cambio en su cultura. Se debe transmitir la importancia del dato a todos los colaboradores para que la perdida de información sea lo menor posible y la explotación pueda realizarse de manera eficiente. En el dataset que se utilizó para este trabajo contamos con información que es obtenida en forma automática pero también datos que son ingresados por los empleados al contactar a los clientes o al darlos de alta en el sistema. Para ello es fundamental que se transmita y explique la importancia del registro correcto de los datos ya que va a ser la materia prima para el desarrollo de los modelos de aprendizaje automático.

En una entidad financiera los datos relacionales de los clientes provienen principalmente de 3 fuentes:

- Operaciones financieras de los clientes (datos estructurados).

- Comentarios en redes sociales (No estructurados).

- Encuestas abiertas (semi estructurados).

El rol de la arquitectura de datos es fundamental para administrar el volumen y la diversidad de datos. La misma debe comenzar el ciclo con un sistema de integración de datos que capture la información de las distintas fuentes y las codifique para poder almacenarlas dentro de un sistema de almacenamiento o Data Lake. Una vez registrada la información se puede comenzar con el análisis exploratorio de datos y la aplicación de

modelos de aprendizaje automático. Por último, dentro de toda arquitectura de datos se utiliza un software en el cual se presentan los reportes y los resultados obtenidos.

A fines prácticos la base que se utilizó para el presente trabajo contiene únicamente datos estructurados, ya que lo que se busca es demostrar los beneficios de la aplicación de modelos de aprendizaje automático en contextos organizacionales. Dicha aplicación brinda a la organización herramientas que permiten mejorar el ratio de conversión de sus campañas, segmentando a los potenciales clientes en categorías dentro de las cuales se establece las condiciones de contacto que mayores probabilidades de éxito tendrá. Aplicar estas técnicas dentro de un marco de metodología ágil brinda la posibilidad de poder retroalimentar el modelo al comparar los resultados pronosticados con la realidad y de esta forma ajustar los parámetros para mejorar la performance. Otro de los beneficios de trabajar con esta metodología es que al poder entregar de forma temprana un MVP (producto mínimo viable) los colaboradores van a poder beneficiarse rápidamente de la aplicación del modelo, lo que los motiva a comprometerse en la correcta utilización de los datos.

Para finalizar el presente apartado es importante hablar de los desafíos que enfrenta la industria financiera en materia de seguridad en el manejo de datos. Como consecuencia de las medidas de prevención del Covid-19 los Bancos trasladaron a gran parte de sus colaboradores a trabajar en forma remota. Al hacerlo en muchos casos las condiciones de seguridad se vieron afectadas por no contar con las mismas medidas de protección que los ordenadores de la empresa. Un relevamiento realizado por la empresa Deloitte[4] informa que entidades financieras como otras organizaciones están sufriendo vulnerabilidades en materia de seguridad de la información, y organismos como el FBI y la Organización Mundial de la Salud alertan sobre las precauciones que deben tomarse.

Apartado 2. Descripción metodológica

2.1. Recopilación de la información

La recopilación de la información para este trabajo se realizó a través de un método conocido como revisión sistemática, el cual es aplicado en muchas investigaciones

[4] https://www2.deloitte.com/es/es/pages/risk/articles/covid-19-gestion-fraude-entidades-financieras.html

científicas. El mismo parte de la búsqueda de información que se adecue a la pregunta o a la hipótesis del trabajo. Para el mismo se recurrió a bibliografía relacionada con la disciplina y el criterio de inclusión se centraba en su relación con el objeto de estudio del trabajo, la gestión de datos en contextos organizacionales. Existe una inmensa cantidad de literatura relacionada al tema principal por lo que el foco se orientó en buscar su relación con modelos de aprendizaje automático y algoritmos de predicción.

A su vez a medida que se fue recolectando información se evaluaron distintos dataset que presentaban las condiciones requeridas, y se terminó optando por el publicado en el sitio de Microsoft Azure[5] ya que cumplía con las condiciones necesarias para ayudar a validar la hipótesis planteada. Se necesitaba un dataset que contenga medios no presenciales para la gestión de campañas y que combinen el uso de elementos clásicos como las llamadas salientes, y también instrumentos más contemporáneos. Lo que se buscaba era hacer incapie en la necesidad de hacer una gestión de campañas personalizada dejando de lado las campañas masivas tradicionales que utilizan gran cantidad de recursos teniendo baja performance.

El siguiente cuadro reporta información brindada por la gerencia de canales del Banco Galicia sobre la cantidad y proporción de ventas de seguros a través de los distintos canales para el primer trimestre del 2019. Como se puede verificar la modalidad presencial a través de las sucursales es sin duda la principal fuente de ganancias para este negocio. El año 2020 cambio los modelos de negocios de las principales empresas del mundo, y obviamente esta empresa no quedo afuera. Al reducirse significativamente la afluencia de personas a las sucursales afecto significativamente los ingresos para este rubro. Como consecuencia se acelera la necesidad de poder optimizar la gestión de campañas de manera no presencial lo que acentúa el propósito del presente trabajo.

Canal	Monto $	Distribución
Sucursales	13.940.339	60%
Online Banking	2.308.382	10%
Telemarketing	4.181.898	18%
Conecta	653.064	3%
UVM	850.238	4%
Call Center	1.234.102	5%
Total	23.168.023	

Distribución de ventas de seguros. Fuente: Banco Galicia S.A

[5] https://gallery.azure.ai/Solution/Campaign-Optimization-with-SQL-Server

2.2. Procesamiento de la información

El surgimiento de la era del Big Data cambió la forma de hacer negocios en todo el mundo. Boyd y Crawford (2012) definen al Big data como un fenómeno cultural, tecnológico y académico que surge gracias al desarrollo de computadoras potentes y algoritmos precisos que permiten analizar grandes volúmenes de datos y obtener un nuevo conocimiento el cual de manera previa era imposible. El mundo de las organizaciones está comenzando a aprovechar los beneficios de dicho fenómeno, el cual puede ser aplicado para todos los procesos que las atraviesan. Grandes volúmenes de datos se generan a diario en las organizaciones y el desafío con el que muchas se encuentran es como utilizarlos para aumentar su rentabilidad. A su vez el almacenamiento representa un costo importante haciendo más relevante centrarse en su optimización. El dilema es ¿cómo utilizar los datos para hacer más eficientes sus procesos?

La minería de datos está transformando la visión estratégica de las organizaciones, siendo fundamental para el desarrollo de los objetivos su correcta implementación. El procesamiento de la información en el presente trabajo se realizó utilizando una combinación de distintos softwares de gestión de datos que permitieron unir las bases que contaban con información relevante para luego poder editarla y explotarla.

2.3. Análisis de la información

El contenido bibliográfico sirvió de guía para poder aplicar las técnicas de gestión de datos de manera eficiente y poder llegar al objetivo deseado. Cabe mencionar que a medida que se fueron aplicando los ajustes en el dataset y obteniendo resultados tanto la óptica del trabajo como la forma de llegar al objetivo fue modificándose. Así como se plantean los beneficios que trae a las organizaciones trabajar en una metodología ágil, el desarrollo de este trabajo también tuvo ciertas características de agilidad. Al correr los modelos y analizar la información obtenida se fueron generando distintos insights que generaron nuevas ópticas para aportar valor a la disciplina academia y organizacional, los cuales no eran el objetivo principal del trabajo. En lugar de dejarlos de lado y continuar con el objetivo principal, se decidió ampliar el alcance para poder hacer un mayor énfasis en la utilidad que trae la implementación de este tipo de trabajos. El objetivo propuesto

era generar un esquema de recomendación que sirva de guía al momento de gestionar las campañas, pero durante el desarrollo surgió la posibilidad de crear un modelo que además permita adaptarse a las necesidades de distintos tipos de organizaciones que cuentan con menor cantidad de recursos y necesitan una base reducida, pero con mayor efectividad para realizar sus campañas.

Apartado 3. Implementación.

3.1 Puesta en producción de un modelo.

El primer paso para el proceso de implementación de modelos de aprendizaje automático es la definición del problema, el cual consiste en optimizar la base de datos de una entidad bancaria, para obtener una mejor performance en la gestión de campañas comerciales. Una vez definido el problema se pasa a la parte de entendimiento de los datos y preparación de estos.

3.1.1 Análisis Exploratorio y preparación de datos.

Es un conjunto de herramientas y técnicas de visualización de datos cuyo fin es poder hacer un análisis preliminar. El mismo genera entendimiento del problema, garantiza que se están haciendo las preguntas correctas, permite visualizar los elementos y realizar nuevos insights. Su objetivo principal es detectar relaciones entre variables, estimar las direcciones y dimensiones de las variables involucradas. A su vez permite detectar errores que pueda contener el dataset, verificar las suposiciones que se consideraban a priori e identificar los modelos que se pueden llegar a aplicar.

El Análisis exploratorio de datos puede realizarse en forma gráfica y no gráfica, así como también puede ser Univariado y multivariado. Su aplicación persigue distintos fines y genera distintos resultados como se verá a continuación.

Univariado y no gráfico

El objetivo es tener una mejor apreciación de como luce la distribución de las variables, analizando de una variable a la vez. Se observan las frecuencias de los datos y la detección de outliers. Gracias a esto se puede verificar que la única columna que presenta valores nulos en el dataset es la de tipo de educación. Al comprobar que los mismos representan

un 1% del total de los valores de su columna se decide omitirlos ya que no afectaran el modelo que se va a utilizar.

Univariado y gráfico

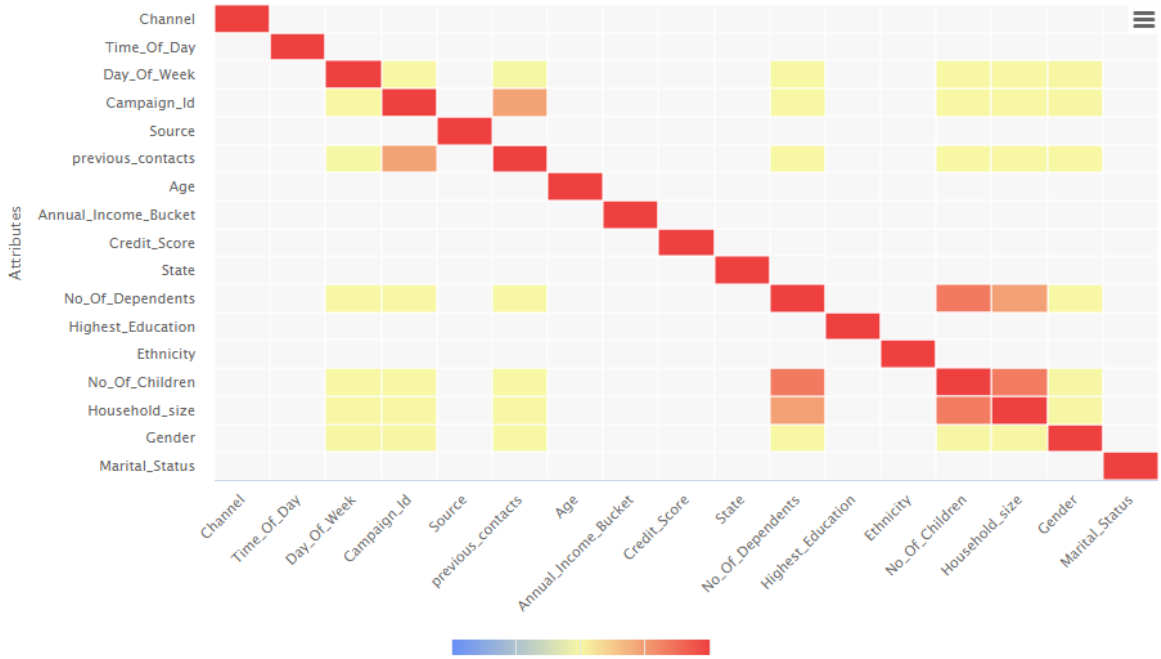
El análisis grafico de este tipo complementa al anterior dando una mirada general de las variables. Es de mucha utilidad para la detección de errores o outliers, los cuales representan valores diferentes o inusuales con respecto a los demás. No se verifican outliers así como tampoco errores en el dataset original.

Multivariado y no gráfico

Generalmente muestran la relación entre varias variables a través, en el presente dataset al ser variables categóricas no encontramos correlaciones significativas que sean de utilidad para el análisis.

Matriz de correlación.

En la siguiente matriz de correlaciones se pueden observar que en la mayoría de los cruces entre variables no existe correlación o si la hay es baja, salvo para las variables de cantidad de hijos, tamaño del hogar y número de dependientes.



Matriz de correlaciones *Fuente: Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)*

Para evitar problemas de correlación se procede a dejar únicamente la variable cantidad de hijos reduciendo también el tamaño del dataset lo que facilita su utilización.

Multivariado Gráfico

De manera gráfica permite comenzar a entender las relaciones entre las variables. Esta etapa es muy importante porque comienza a dar una idea de cuáles son los resultados que se pueden obtener al momento de aplicar los modelos. En el siguiente gráfico se puede visualizar la relación entre los contactos previos al mismo cliente y las ventas realizadas:

Cantidad de ventas según contactos previos Fuente: Elaboración propia

A partir de la visualización se descubre que todas las ventas realizadas fueron a clientes que se los contacto cuatro o más veces. Esto brinda dos conclusiones. Por un lado, que los clientes con mayor cantidad de contactos previos es más probable que tomen el seguro. Por otro lado, a nivel organizacional se puede plantear como desafío entender porque a los clientes que fueron contactados menos de cuatro veces no se logró vender el seguro y tomar medidas para mejorar esa variable.

A su vez dentro del análisis y exploración de datos es posible aplicar medidas para conocer con mayor profundidad los resultados. A través de la herramienta Power Bi se obtuvo el porcentaje de conversión de acuerdo con la variable contactos previos.

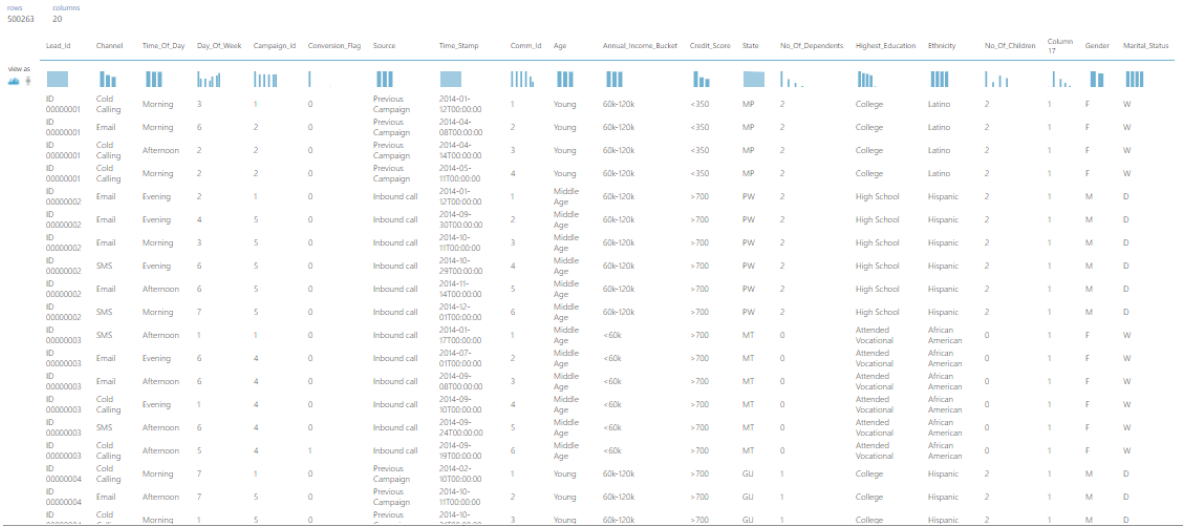
Porcentaje de ventas según contactos previo. *Fuente: Elaboración propia*

Este gráfico demuestra que a medida que la cantidad de contactos que se realiza aumenta la posibilidad de éxito también lo hace. Como se planteó anteriormente uno de los beneficios principales del machine learning es la capacidad de retroalimentarse a través de su uso. Con la aplicación de modelos como los de este trabajo se busca potenciar la misma base sobre la cual se trabaja, entonces al actualizarla con las mejoras aplicadas se espera que los resultados de los nuevos dataset permitan generar mejores modelos que se ajusten a partir de su aplicación.

El Análisis exploratorio de datos es fundamental porque brinda una visión general necesaria para editar los parámetros de las distintas variables y asegurarse que sean compatibles con los algoritmos seleccionados para obtener resultados óptimos. A partir del mismo se tomó la decisión de que al momento de correr el modelo se van a modificar las variables días de la semana y tipo de campaña que estaban representadas por números y cambiarlas por letras para que no afecten la fórmula de la regresión. A la vez se decidió eliminar la columna de la fecha ya que afecta el desempeño del modelo al generar un sobreajuste en la predicción.

3.1.2 Puesta en marcha

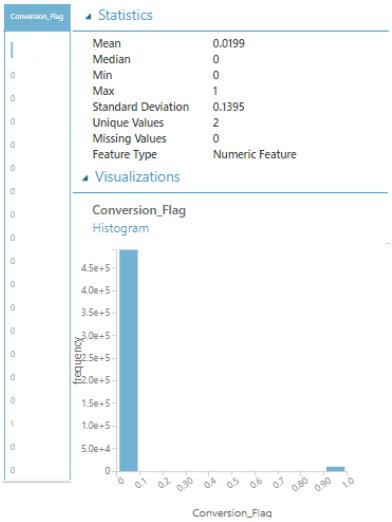
La herramienta de Microsoft Azure Machine Learning permite realizar una exploración inicial al momento de cargar el dataset y conocer las frecuencias y tipos de variables. El dataset original consta de 500.263 registros y 20 variables las cuales incluyen datos del momento en que se realizó el contacto, el medio, la cantidad de contactos previos y luego datos demográficos como ingresos, edad, tipo de educación, ocupación, estado civil y cantidad de familiares.



Lead_Id	Channel	Time_Of_Day	Day_Of_Week	Campaign_Id	Conversion_Flag	Source	Time_Stamp	Comm_Id	Age	Annual_Income_Bucket	Credit_Score	State	No_Of_Dependents	Highest_Education	Ethnicity	No_Of_Children	Column 17	Gender	Marital_Status
ID 00000001	Cold Calling	Morning	3	1	0	Previous Campaign	2014-01-12T00:00:00	1	Young	60k-120k	<350	MP	2	College	Latino	2	1	F	W
ID 00000001	Email	Morning	6	2	0	Previous Campaign	2014-04-08T00:00:00	2	Young	60k-120k	<350	MP	2	College	Latino	2	1	F	W
ID 00000001	Cold Calling	Afternoon	2	2	0	Previous Campaign	2014-04-14T00:00:00	3	Young	60k-120k	<350	MP	2	College	Latino	2	1	F	W
ID 00000001	Cold Calling	Morning	2	2	0	Previous Campaign	2014-05-11T00:00:00	4	Young	60k-120k	<350	MP	2	College	Latino	2	1	F	W
ID 00000002	Email	Evening	2	1	0	Inbound call	2014-01-12T00:00:00	1	Middle Age	60k-120k	>700	PW	2	High School	Hispanic	2	1	M	D
ID 00000002	Email	Evening	4	5	0	Inbound call	2014-09-30T00:00:00	2	Middle Age	60k-120k	>700	PW	2	High School	Hispanic	2	1	M	D
ID 00000002	Email	Morning	3	5	0	Inbound call	2014-10-11T00:00:00	3	Middle Age	60k-120k	>700	PW	2	High School	Hispanic	2	1	M	D
ID 00000002	SMS	Evening	6	5	0	Inbound call	2014-10-29T00:00:00	4	Middle Age	60k-120k	>700	PW	2	High School	Hispanic	2	1	M	D
ID 00000002	Email	Afternoon	6	5	0	Inbound call	2014-11-14T00:00:00	5	Middle Age	60k-120k	>700	PW	2	High School	Hispanic	2	1	M	D
ID 00000002	SMS	Morning	7	5	0	Inbound call	2014-12-01T00:00:00	6	Middle Age	60k-120k	>700	PW	2	High School	Hispanic	2	1	M	D
ID 00000003	SMS	Afternoon	1	1	0	Inbound call	2014-01-17T00:00:00	1	Middle Age	<60k	>700	MT	0	Attended Vocational	African American	0	1	F	W
ID 00000003	Email	Evening	6	4	0	Inbound call	2014-07-01T00:00:00	2	Middle Age	<60k	>700	MT	0	Attended Vocational	African American	0	1	F	W
ID 00000003	Email	Afternoon	6	4	0	Inbound call	2014-09-08T00:00:00	3	Middle Age	<60k	>700	MT	0	Attended Vocational	African American	0	1	F	W
ID 00000003	Cold Calling	Evening	1	4	0	Inbound call	2014-09-10T00:00:00	4	Middle Age	<60k	>700	MT	0	Attended Vocational	African American	0	1	F	W
ID 00000003	SMS	Afternoon	6	4	0	Inbound call	2014-09-24T00:00:00	5	Middle Age	<60k	>700	MT	0	Attended Vocational	African American	0	1	F	W
ID 00000003	Cold Calling	Afternoon	5	4	1	Inbound call	2014-09-19T00:00:00	6	Middle Age	<60k	>700	MT	0	Attended Vocational	African American	0	1	F	W
ID 00000004	Cold Calling	Morning	7	1	0	Previous Campaign	2014-02-10T00:00:00	1	Young	60k-120k	>700	GU	1	College	Hispanic	2	1	M	D
ID 00000004	Email	Afternoon	7	5	0	Previous Campaign	2014-10-11T00:00:00	2	Young	60k-120k	>700	GU	1	College	Hispanic	2	1	M	D
ID	Cold	Morning	1	5	0	Previous	2014-10-	3	Young	60k-120k	>700	GU	1	College	Hispanic	2	1	M	D

Visualización del dataset original a través del Software de Microsoft Azure Machine Learning. Fuente: Elaboración propia

Del total de contactos se lograron 9939 ventas, el porcentaje de conversión es de un 1.99% y el objetivo del presente trabajo es optimizar la forma en que se realizan las campañas para obtener mejores resultados y hacer un uso más eficiente de los recursos organizacionales a partir de los datos disponibles. A continuación, se puede visualizar la salida del Azure ML seleccionando la variable a predecir a través de la función visualize:



Porcentaje de conversión de Campañas. Fuente: Elaboración propia

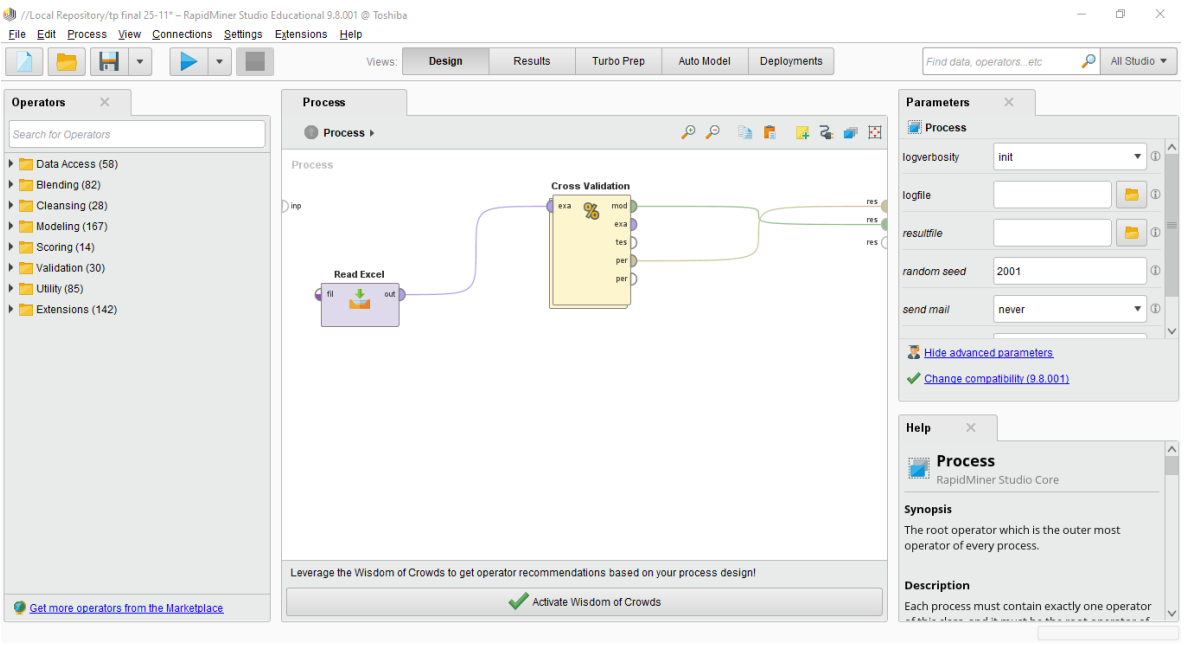
En el siguiente enlace se puede acceder a un video donde se explica y amplia la aplicación de las técnicas de análisis y exploración de datos a través de la herramienta Microsoft Azure Machine Learning, R studio y Power Bi: *https://www.youtube.com/watch?v=fm0t279VSPQ&feature=youtu.be*

3.1.3 Selección de algoritmo y entrenamiento del modelo.

Deshpande y Kotu en Predictive Analytics and Data Mining (2014), sostienen que hay dos clases principales de técnicas de análisis predictivo: las que evolucionaron a partir de estadísticas (como las regresiones) y las que surgieron de una combinación de estadísticas, ciencia y matemáticas (como árboles de clasificación). La elección del modelo depende de los tipos de variables con las que se cuenten. El problema que se plantea en este trabajo es predecir una variable categórica binomial (Si/No), como resultado de los contactos realizados. Para su predicción el dataset contiene distintos atributos que se comportan como variables independientes y se tratará de explicar el resultado del llamado a través de ellas. La técnica que se va a utilizar es regresión logística, con la cual se espera obtener los coeficientes que expliquen la variable dependiente a partir de las variables independientes significativas, aquellas cuyo p valor sea menor a 0.05.

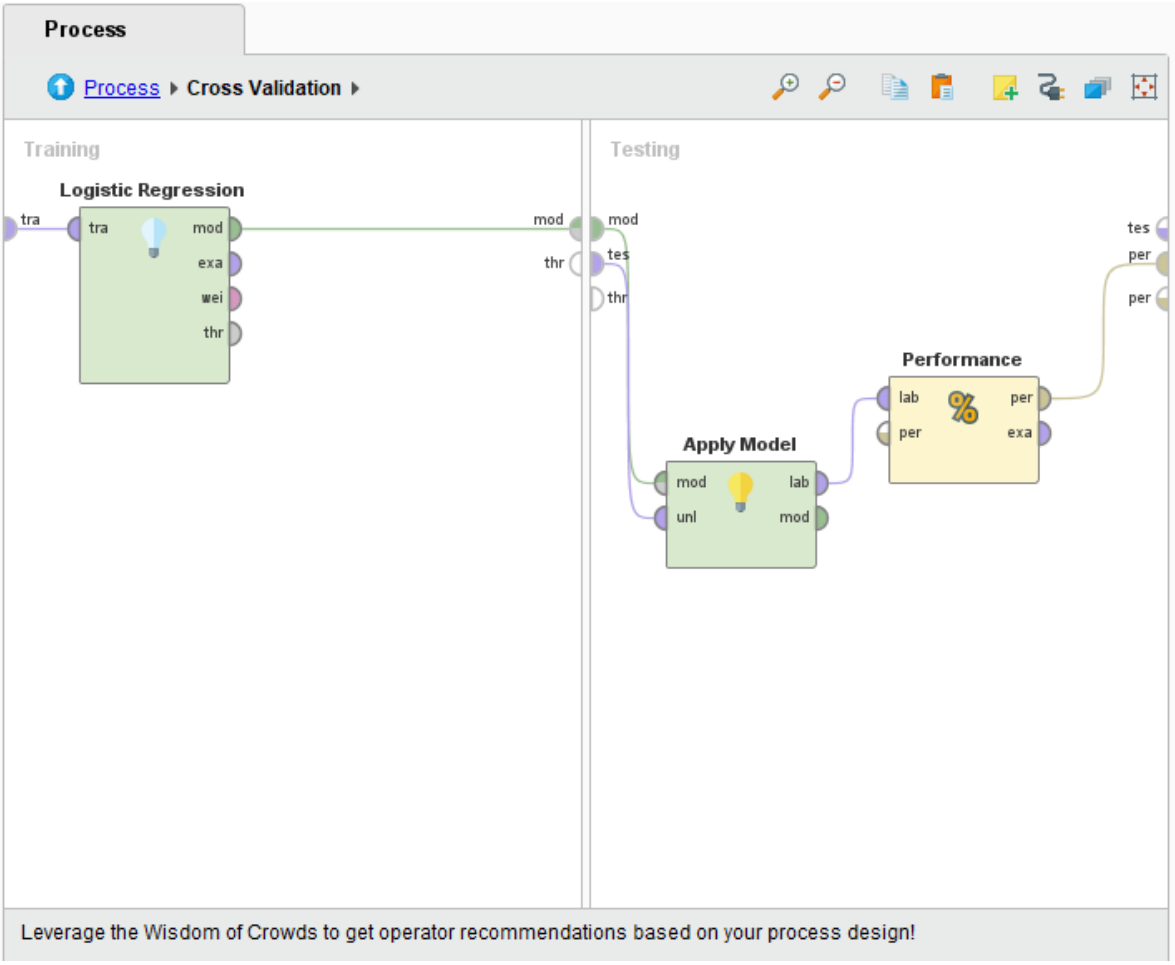
Además, utilizando el Software Rapid Miner se contará con la matriz de confusión la cual indicara el porcentaje de accuracy del modelo, si bien no es el objetivo principal del trabajo, sirve como validación de que se está utilizando un algoritmo correcto. Para obtener dicha matriz es necesario un método de validación que cruce las predicciones realizadas por el modelo contra la realidad. Para ello se deben separar los datos en entrenamiento y prueba siendo de los primeros desde donde se extraerá la información de las variables para tratar de predecir los resultados en el set de prueba. El tipo de validación que se utilizó en este trabajo es Cross Validation o validación Cruzada. Este operador tiene dos subprocesos: un subproceso de formación y un subproceso de prueba. La base original es partida en k bloques de igual tamaño. Luego se toma un bloque para usarlo de testeo y el resto se deja para entrenamiento. Esto se repite un número determinado de veces siempre utilizando un bloque distinto como testeo y de allí se obtiene la validación cruzada, reduciendo el riesgo de sobreajuste del modelo. (© RapidMiner GmbH, 2020).

RapidMiner es una herramienta intuitiva y de uso sencillo ya que su interfaz permite unir los distintos procesos a través de conectores sin tener la necesidad de conocimientos previos en lenguajes de programación. El primer paso es cargar el dataset y seleccionar los hiper parámetros según correspondan, es necesario aclara que la variable a predecir es binomial para que el operador de regresión logística funcione. Luego se conecta el dataset a la entrada de Cross validation.



Conexión dataset a Cross Validation en Rapid Miner Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Una vez que están los datos cargados en paso siguiente se deben ajustar las métricas para el operador de regresión logística, luego se utiliza Apply model que toma la porción de entrenamiento y de prueba y lo valida con el indicador de performance seleccionado.



Subproceso Cross Validation. Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Como resultado se obtiene la matriz de confusión que en el lado derecho indica el porcentaje de acierto para cada tipo de resultado, además de la performance global del modelo.

accuracy: 98.05% +/- 0.02% (micro average: 98.05%)

	true 0	true 1	class precision
pred. 0	489934	9358	98.13%
pred. 1	390	581	59.84%
class recall	99.92%	5.85%	

Matriz de confusión RapidMiner. Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

El Baseline es del 98,01%, con el modelo utilizado se lo pudo superar mínimamente ya que dataset está sumamente desbalanceado. Si bien el objetivo del trabajo no es superar el Baseline, el resultado indica que el modelo funciona. Pensando en su aplicación para el negocio si la organización decide solamente contactar a aquellos cuya predicción va a ser positiva la precisión es del 60% logrando 581 ventas por sobre 971 contactos. La utilidad de este resultado va a depender del tipo de organización y los recursos con los que cuente, ya que, si bien se asegura un mayor porcentaje de conversión, se están dejando por fuera 9358 ventas posibles.

3.2 Visualizaciones de resultados

3.2.1 Primera parte

Adicionalmente al performance el software RapidMiner entrega el resultado del algoritmo utilizado, que en el caso de regresión logística son los coeficientes de la ecuación que se empleó para predecir los resultados. Dicha información es lo que interesa a fines de poder comenzar a comprender cuales son las variables independientes que mayor incidencia tienen en la variable respuesta. A continuación, se presenta la tabla con los principales resultados de la regresión con incidencia positiva en la conversión de campañas.

A partir de estos resultados se pudo conocer cuáles son las variables que aseguraban mejores resultados orientando la dirección hacia el objetivo del trabajo.

Attribute	Coefficient ↓	Std. Coefficient	Std. Error	z-Value	p-Value
Credit_Score.350-700	1.889	1.889	0.027	69.518	0
previous_contacts	1.082	1.652	0.010	107.814	0
Highest_Education.Hi...	1.003	1.003	0.033	30.000	0
Highest_Education.Att...	0.657	0.657	0.034	19.270	0
Highest_Education.Na	0.602	0.602	0.109	5.519	0.000
Credit_Score.>700	0.501	0.501	0.033	15.423	0
No_Of_Children	0.241	0.288	0.014	17.220	0
Age.Senior Citizen	0.204	0.204	0.028	7.165	0.000
State.PR	0.169	0.169	0.112	1.503	0.133
Highest_Education.Gr...	0.153	0.153	0.036	4.241	0.000
State.CT	0.137	0.137	0.116	1.183	0.237
State.SD	0.128	0.128	0.114	1.123	0.261
State.CO	0.127	0.127	0.115	1.106	0.269
Age.Middle Age	0.112	0.112	0.028	3.998	0.000
State.NE	0.112	0.112	0.115	0.980	0.327

Logistic Regression Model RapidMiner 1 Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Realizando un filtro a solo contactos cuyo Score crediticio está en el rango entre 350 y 700 (se podría definir como Score medio) el porcentaje de conversión de la campaña sube a un 4,94% logrando 5919 ventas lo que representa casi un 60% del total. Continuando con las variables principales las personas cuyo nivel educativo es High School son las que mayor propensión tienen, entonces marcando dicha variable el porcentaje de acierto aumenta a 7,35%.

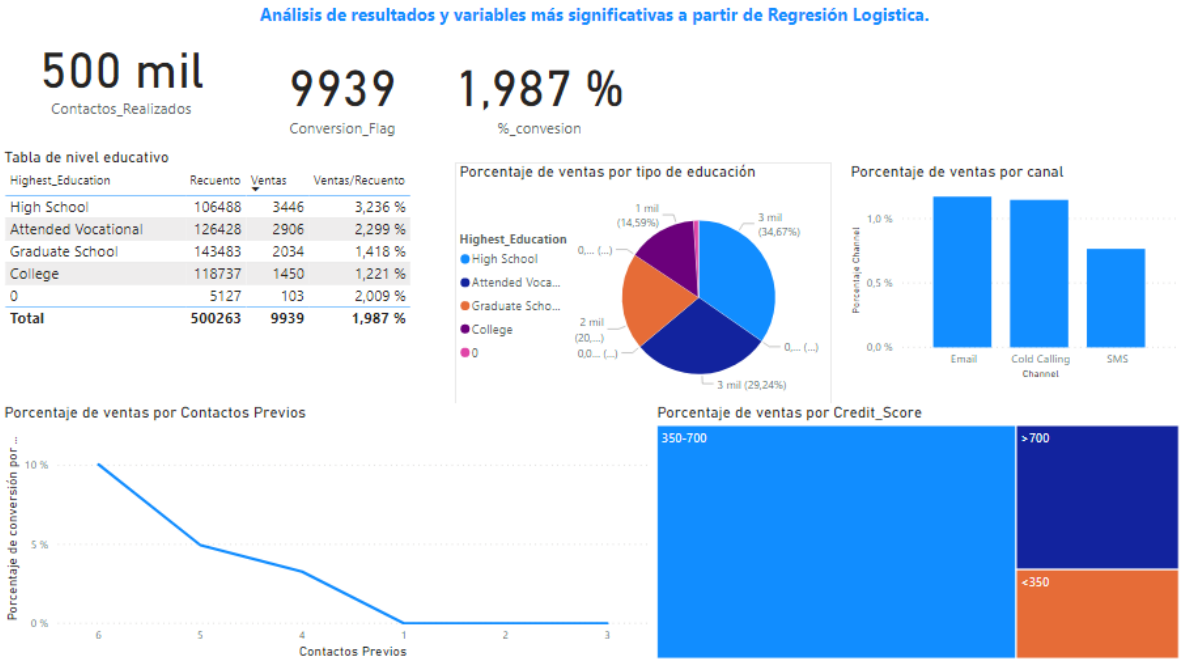
Como se verificó al momento del Análisis exploratorio de datos la variable contactos previos es importante ya que no hubo conversiones para los valores menores a 4, el coeficiente de la regresión lo confirma. Además de tener en cuenta las variables con mayor incidencia positiva se tuvo en cuenta las negativas en el cuadro que se presenta a continuación:

Attribute	Coefficient ↑	Std. Coefficient	Std. Error	z-Value	p-Value
Intercept	-9.347	-6.102	0.118	-79.363	0
Channel.SMS	-0.464	-0.464	0.027	-16.895	0
Gender.M	-0.376	-0.376	0.022	-17.287	0
State.MT	-0.348	-0.348	0.125	-2.788	0.005
Time_Of_Day.Afternoon	-0.326	-0.326	0.028	-11.781	0
Annual_Income_Buck...	-0.197	-0.197	0.029	-6.875	0.000
State.ND	-0.185	-0.185	0.121	-1.531	0.126
State.MN	-0.165	-0.165	0.117	-1.409	0.159
State.KY	-0.162	-0.162	0.121	-1.333	0.182
State.MA	-0.156	-0.156	0.120	-1.298	0.194
State.OH	-0.153	-0.153	0.120	-1.284	0.199
State.AR	-0.133	-0.133	0.119	-1.114	0.265
State.OR	-0.113	-0.113	0.122	-0.928	0.354
State.LA	-0.112	-0.112	0.119	-0.945	0.345
Day_Of_Week	-0.103	-0.227	0.006	-17.123	0

Logistic Regression Model RapidMiner 2 Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

En base a lo que representa la salida para los coeficientes negativos las variables no son tan significativas como la positiva, pero comprobando en forma iterativa y en base a la premisa establecida en el planteo del objetivo se tomó la variable SMS la cual graficando su porcentaje de conversión se comprueba como el más bajo de las 3.

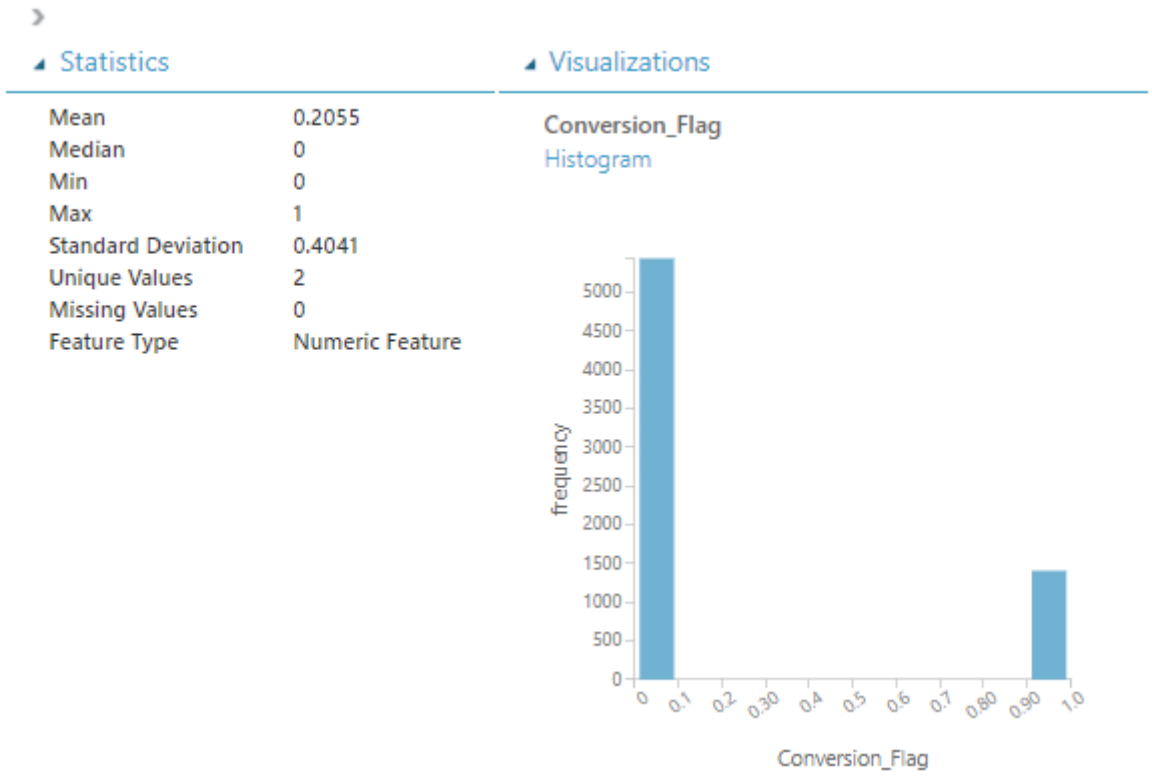
Se utilizo la herramienta Power Bi para validar los resultados extraídos del modelo de regresión y representar en forma gráfica la performance de las variables más significativas:



Representación Gráfica en Power Bi *Fuente: Elaboración propia*

Como se puede apreciar en la tabla y en el gráfico la variable High Schol representa el mayor porcentaje de conversión para el nivel educativo. Lo mismo se ve para el canal de ventas que tiene un porcentaje menor a los otros dos medios. La curva de contactos previos desciende a medida que se reducen la cantidad de contactos previos. Por último, a simple vista se puede ver que la variable Score medio es significativamente mayor al resto entendiendo el motivo por el cual su coeficiente era el más alto. Con la idea de poner a prueba la información obtenida, se elaboró un dataset seleccionando solamente las variables mencionadas como las más significativas.

A continuación, se grafican los resultados de este nuevo Dataset:



Porcentaje de conversión de Campañas con dataset filtrado *Fuente: Elaboración propia*

Sobre la base original se seleccionaron solamente personas con Score medio, nivel de educativo High School, que anteriormente recibieron cuatro o más contactos y el medio de contacto fue uno distinto a SMS. Al realizarlo se obtuvo un porcentaje de conversión del 20,55% cuando el original era menor al 2%.

Este resultado es sumamente relevante para las organizaciones ya que permite conocer los perfiles con mayor propensión a contratar el seguro y los medios más adecuados. A su vez para organizaciones que cuentan con una menor estructura que un banco y que realizar 500 mil contactos resulta inviable, aplicar modelos de aprendizaje automático para conocer en que perfiles centrarse puede resultar un diferencial para su negocio. A continuación, se presenta la matiz de confusión para el dataset reducido:

accuracy: 80.82% +/- 1.16% (micro average: 80.82%)

	true 0	true 1	class precision
pred. 0	5301	1178	81.82%
pred. 1	134	228	62.98%
class recall	97.53%	16.22%	

Matriz de confusión para dataset filtrado Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Como se puede observar se reducen significativamente los contactos negativos y el modelo tiene un acierto del 63% sobre los contactos que se predicen como positivos. Si se aplica este modelo solamente con 362 contactos se pueden realizar 228 ventas. Al igual que sucede con el dataset original. Al estar desbalanceado el dataset se siguen perdiendo gran cantidad de ventas posibles.

Este tipo de análisis también permite pensar en otro tipo de optimización propuesto. Una organización si desea puede centrar sus esfuerzos en un solo segmento del mercado y del total de su base enfocarse en aquellos que el modelo demuestra mejor propensión. Por ejemplo, en base a la información obtenida en los párrafos anteriores se podría armar una campaña de marketing para aquellos clientes que encajen en el perfil de score medio y estudios secundarios. Por otro lado, si se desea ganar participación en distintos mercados, se puede utilizar para enfocar los recursos en los segmentos que demuestran peor performance y analizar qué medidas tomar para mejorar el productos o servicio ofrecido a dicho público.

3.2.2 Segunda parte

Continuando con el objetivo principal del trabajo lo que se busca es conocer cuál es la mejor forma de contactar a los clientes en base a su perfil. Si bien los resultados obtenidos hasta ahora brindan un dataset reducido con los perfiles más propensos a contratar el seguro, no responde a la premisa inicial. Un Banco cuenta con la capacidad de poder contactar a cientos de miles de personas en el año entonces el solo hecho de contar con un dataset reducido de personas con mayor propensión no resuelve el problema de optimización.

Lo que se busca es que al momento de realizar el contacto se conozca en base al perfil del cliente, cuáles son las mejores condiciones para hacerlo. Existen diversas formas de segmentar a los clientes y de establecer las condiciones, pero en base a los coeficientes obtenidos en el modelo de regresión se va a segmentar el perfil en base al grupo etario, y las condiciones harán referencia al tipo de contacto y al momento del día. A continuación, se presenta la tabla que obtenida a partir del modelo que se desarrollo que da respuesta a la premisa inicial.

Recomendación de Canal y momento del día para grupo etario					
Channel	**Age**	**Time_Of_Day**	**Contactos**	**Ventas**	**% de conversión**
Cold Calling	Senior Citizen	Morning	5919	933	15,763%
SMS	Young	Evening	11305	1063	9,403%
Email	Young	Evening	6654	558	8,386%
Email	Senior Citizen	Morning	8985	696	7,746%
Cold Calling	Middle Age	Evening	11398	664	5,826%
Cold Calling	Middle Age	Morning	11095	563	5,074%
Cold Calling	Middle Age	Afternoon	11240	565	5,027%
SMS	Young	Afternoon	10691	519	4,855%
Email	Young	Afternoon	6408	249	3,886%
Cold Calling	Senior Citizen	Afternoon	10248	315	3,074%
Cold Calling	Senior Citizen	Evening	10342	285	2,756%
SMS	Senior Citizen	Morning	20494	464	2,264%
SMS	Young	Morning	20880	355	1,700%
Email	Middle Age	Afternoon	21678	348	1,605%
Email	Middle Age	Morning	21484	339	1,578%
Email	Middle Age	Evening	22158	325	1,467%
Email	Young	Morning	12244	177	1,446%
Email	Senior Citizen	Afternoon	16951	231	1,363%
Email	Senior Citizen	Evening	16771	198	1,181%
SMS	Middle Age	Afternoon	21311	178	0,835%
SMS	Middle Age	Evening	22085	164	0,743%
SMS	Middle Age	Morning	21352	157	0,735%
Cold Calling	Young	Evening	24542	163	0,664%
SMS	Senior Citizen	Afternoon	40427	158	0,391%
SMS	Senior Citizen	Evening	40044	133	0,332%
Cold Calling	Young	Afternoon	25063	69	0,275%
Cold Calling	Young	Morning	48494	70	0,144%

Ranking de recomendación *Fuente: Elaboración propia*

Finalmente se obtiene el sistema de recomendación a partir del tipo de contacto y rango de edad. Esta tabla brinda información útil al momento en que se decide realizar un contacto ya que se conoce cuál puede ser el método más eficaz según el perfil del cliente. Por ejemplo, si estamos ante un caso de un ciudadano mayor el medio por excelencia para contactarlo va a ser una llamada por la mañana, en cambio sí es a un joven va a ser a través de un email o sms por la noche. En términos estadísticos se cuenta con el respaldo empírico, pero lo que no deja de ser importante es el análisis más allá de los números y en este caso se puede entender dese el punto de vista sociológico. Los ciudadanos mayores tienden a tener una mayor desconfianza o menor uso de las nuevas tecnologías entonces es entendible que a una persona joven sea más efectivo comunicarlo a través de un email, sms o whatsapp, que en Argentina es el medio de comunicación por antonomasia. En oposición los jóvenes cada vez usan menos el teléfono para llamados y así se verifica en la tabla que las ventas a través de llamados telefónicos para ese segmento ocupan los últimos lugares.

En el punto siguiente se explicará la metodología que se recomienda para aplicar este proyecto. Lo importante es entender que los de aprendizaje automático obtienen su mejor resultado a medida que entran en juego con el ambiente externo. Aplicar modelos de aprendizaje automáticos en organizaciones que trabajan bajo la metodología Agile permite una retroalimentación que brinda mejores resultados a medida que se va desarrollando. Al comenzar a realizar los contactos siguiendo los aprendizajes obtenidos por el modelo va a brindar mayores datos orientados a mejorar la precisión del modelo y a la corrección de los errores que se puedan haber incurrido o que permitan corregir desvíos producidos por cambios propios del ambiente dinámico en el que se desenvuelve.

3.3 Metodologías ágiles para la implementación de proyectos de aprendizaje automático.

3.3.1 Problemática organizacional.

La industria financiera se caracterizó históricamente por una estructura tradicional, la cual no requirió de innovaciones en materia organizacional para tener obtener ganancias para sus accionistas. El surgimiento de las Fintech fue disruptivo para un sector bancario, el cual esta ese momento no había tenido en la historia un nuevo actor que funcione como

competencia indirecta y amenace su participación en el mercado. Los Bancos habituados a una escaza competencia y a alianzas estratégicas que les permitían monopolizar el mercado se encontraron con una amenaza, nativa digital y dispuesta a poner sus servicios al alcance de todos los consumidores.

Adicionalmente el surgimiento del fenómeno del Big Data transformo el desarrollo de las organizaciones, generando oportunidades multimillonarias para las que aprovechen sus beneficios, y llevando a la quiebra muchas que no. Dentro de este marco el sector financiero comenzó a transformar sus estructuras, en parte para poder hacer frente a la aparición de nuevos competidores y a su vez para poder explotar los beneficios que los nuevos desarrollos tecnológicos podrían dar al negocio.

El Banco Galicia comenzó en el año 2017 un proceso de transformación Digital, el cual hoy se materializa en la búsqueda de convertirse en una organización Data Driven. La dirección estratégica de la organización se propone la meta de la omnicanalidad de sus procesos, tanto para colaboradores como para clientes. Aplicar un cambio de este estilo en una organización con más de seis mil empleados no es fácil y requiere de un cambio en la cultura organizacional que motive a los colaboradores a atravesar ese camino.

El desafío principal consiste en la explotación de los datos obtenidos a través de sus distintas fuentes para potenciar el funcionamiento del negocio y desarrollar servicios que se adapten a las necesidades de los clientes. A su vez esta información correctamente utilizada permite detectar dolores de los clientes y resolverlos en forma temprana.

En una entidad financiera los datos relacionales de los clientes provienen principalmente de 3 fuentes:

- Operaciones financieras de los clientes (datos estructurados).

- Comentarios en redes sociales (No estructurados).

- Encuestas abiertas (semi estructurados).

El rol de la arquitectura de datos es fundamental para administrar el volumen y la diversidad de datos. La misma debe comenzar el ciclo con un sistema de integración de datos que capture la información de las distintas fuentes y las codifique para poder

almacenarlas dentro de un sistema de almacenamiento o Data Lake. Una vez registrada la información se puede comenzar con el análisis exploratorio de datos y la aplicación de modelos de aprendizaje automático. Por último, dentro de toda arquitectura de datos se utiliza un software en el cual se presentan los reportes y los resultados obtenidos.

A fines prácticos la base que se utilizó para el presente trabajo contiene únicamente datos estructurados, ya que lo que se busca es demostrar los beneficios de la aplicación de modelos de aprendizaje automático en contextos organizacionales.

3.3.2 Metodología Agile.

Uno de los pilares de la transformación digital es el desarrollo de metodologías agiles, para poder aportar valor de manera continua, trabajar colaborativamente y entregar valor al cliente. El mundo está cambiando rápidamente, al igual que las expectativas de los clientes. Las organizaciones también necesitan transformarse y la agilidad juega un rol clave en este cambio.

La metodología Agile surge como respuesta ante la creciente velocidad en los avances tecnológicos y los cambios culturales que traen aparentados. La estructura tradicional de las organizaciones se fue adaptando a través del siglo XX a los avances tecnológicos mediante distintas corrientes que modificaban la estructura tradicional jerárquica y centraban el foco en los procesos. La teoría de sistemas transformo la visión estratégica de la organización dejando de entenderla como áreas independientes y pasando a integrar un modelo de red en el cual todos sus sectores debían trabajar en forma alineada en búsqueda de una misma misión guiada por su visión del futuro deseado.

Jose Gilli (2008) plantea “La concepción de la organización como un sistema complejo, resultado de múltiples interacciones entre diferentes componentes, mediante el diseño procura adaptarse a los requerimientos de su medioambiente” (P.22). Ante entornos cada vez más cambiantes las organizaciones necesitan herramientas flexibles que permitan adaptarse rápidamente a las demandas del mercado.

Dentro de este marco llega el manifiesto Agile como una guía de principios para elaborar estructuras organizacionales centradas en el usuario, la respuesta ante el cambio, la creación de ambientes colaborativos y en la rápida entrega de valor al cliente interno o

externo. Las entidades financieras reciben grandes volúmenes de datos en forma diaria, por lo que resulta útil contar con una estructura organizacional que permita la retroalimentación y aprendizaje del modelo a partir de la interacción de sus resultados con la realidad.

3.3.3 Fundamentos del Scrum

Scrum es un marco de trabajo que promueve la innovación y permite que equipos interdisciplinarios y autónomos entreguen resultados de alta calidad en tiempos cortos. El mismo permite entregas parciales que generan rápidamente valor al cliente. Esto reduce el riesgo al ser flexible y dar posibilidad a readaptar los procesos al detectar una falla o un cambio en las condiciones del entorno.

A continuación, se describe como a través del Scrum, uno de los marcos más utilizados a nivel mundial para metodologías agiles, se establece el esquema para el desarrollo de un proceso de aprendizaje automático.

3.3.3 Roles

Scrum propone tres roles en el equipo de trabajo, los cuales cuentan con funcione, objetivos y responsabilidades previamente establecidas. A su vez requiere de constante comunicación entre los roles para asegurar el cumplimiento de la meta perseguida.

Scrum Master

Coach, responsable de potenciar la colaboración y el entendimiento de todo el equipo. Para el Product owner y para el equipo es un facilitador. Debe asegurara que el equipo trabaje ajustándose a las practicas Scrum.

Product owner

Estratega del producto, responsable de que se consiga el impacto esperado. Es el nexo entre con los stakeholders. Su función para el proyecto de optimización consiste en tomar las decisiones sobre que propuestas van a ser trasladadas al equipo a través del Product Backlog. Esto último es una lista ordenada de todo lo que podría ser necesario para el producto.

En este proyecto la interacción con el ambiente externo es muy importante ya que va a ser la retroalimentación lo que permita la mejora del modelo. El equipo entrega el MVP el cual es probado y el PO va a ser quien reciba los resultados y analice cuales son las acciones prioritarias que se van a tomar.

Equipo multidisciplinario

El equipo de desarrollo es el encargado de construir el producto o servicio, la pieza de valor. Cuenta con todas las habilidades necesarias para llevar adelante el modelo. Son responsables de definir cómo harán el trabajo. Debe estar conformado por no más de 6 personas incluyendo analistas, data scientist y un desarrollador Business Intelligence. Los mismos deben tener un perfil colaborativo y se debe compartir el conocimiento para evitar su centralización.

3.3.4 Eventos Scrum

Son instancias predefinidas con el fin de crear un marco en donde se minimicen las reuniones no definidas en el Scrum.

El **Sprint** es un ciclo de trabajo de 4 a 6 semanas durante las cuales el equipo desarrolla el modelo entregable. Durante el Sprint se trabaja de acuerdo con el **Planning** que es un evento en el cual se traza la hoja de ruta. La **Daily** es un espacio abierto de comunicación donde se presentan los avances, los impedimentos que surgieron y la planificación de las próximas 24 horas. Por último, el **Sprint Review** es el momento de revisión donde se evalúa la efectividad del modelo entregado y se adapta el **Product Backlog** a las nuevas condiciones.

Conclusión

La realidad del mundo en que vivimos está cambiando rápidamente, al igual que las expectativas de los clientes y de la sociedad en su conjunto. Las organizaciones también necesitan transformarse y la agilidad juega un rol clave en este cambio. Con el presente trabajo se buscó explicar los beneficios de implementar modelos de aprendizaje automático dentro de un marco de agilidad organizacional con el fin de optimizar los procesos y hacer un uso más eficiente de los recursos. Se llego a la conclusión de que las organizaciones no pueden seguir gestionando de igual manera sus procesos, sino que

deben aprovechar las capacidades de la gestión de datos y los algoritmos predictivos para hacer más eficiente y rentable el trabajo.

Con el modelo desarrollado se buscaba demostrar cómo realizar un modelo que permita optimizar la gestión de las campañas, y brindar un marco de trabajo para que distintos tipos de organizaciones puedan utilizarlo según su realidad y disponibilidad de recursos. A la vez se pueden utilizar estos modelos para buscar implementar mejoras en los productos o servicios ofrecidos, gracias a la información adquirida durante el proceso. El foco en este trabajo estuvo en ofrecer un servicio existente a los clientes dependiendo su perfil, pero este aprendizaje también puede ser utilizado para desarrollar mejoras en lo que se está ofreciendo al mercado analizando las necesidades de cada perfil de consumidor en base al feedback obtenido.

Se puede concluir remarcando que se obtuvieron dos resultados principales. En una primera parte se realizó el análisis de las variables y la preparación de los datos junto con la aplicación del modelo de aprendizaje automático a través de la aplicación del operador regresión logística en la herramienta RapidMiner. Como resultado se logró un modelo que permite conocer aquellos perfiles con mayor propensión a contratar el seguro, un recurso que en un contexto organizacional sumamente competitivo como el de la industria financiera actual, resulta de gran utilidad ya que se pueden centrar principalmente los esfuerzos en ellos. En una segunda parte se persiguió el objetivo principal del trabajo que consistía en elaborar un ranking de recomendación que indique en base al perfil del cliente a través de que medio y en que momento del día contactarlo. Con este resultado el banco cuenta con una guía para la planificación y el desarrollo de sus campañas comerciales, asegurándose un mayor porcentaje de conversión.

Se explicaron además los beneficios de trabajar en una metodología agile, tomando el marco metodológico scrum, uno de los más utilizados en el ambiente empresarial. En comparación con la estructura de planificación de procesos tradicional, agile permite realizar ajustes a medida que se va desarrollando el proceso y poniendo a prueba en el corto plazo los resultados a través de los MVP. Esto brinda dos beneficios, por un lado, la entrega de valor temprana al cliente, y por otro al recibir retroalimentación constante durante el proceso, el objetivo se adapta a los cambios en las condiciones del entorno, entonces el resultado final va a ser mejor. Este ciclo de retroalimentación es lo que

permite potenciar al máximo las capacidades del Machine Learning. Para este trabajo se tomó solamente uno de los casos para los que se pueden utilizar estas herramientas computacionales, que no es algo que se espera implementar en el futuro, sino que ya está sucediendo en las principales organizaciones del mundo, simplemente que su alcance es tan amplio y dinámico que se requieren de mucha inversión para poder aplicarlos.

Bibliografía

- A. McAfee, E. Brynjolfsson (2012), Harvard Business Review, USA.
- Artículo sobre banco Galicia Fernando Raverta:
- Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA. London: Routledge.
- Bustos, J. C. (2016). Big data and Big GAFA. Thoughts on the data economy.
- F. Provost, T. Fawcett (2013), Data Science for Business, USA.
- Gilli, J. Jose, Tartabini, M. Amanda, (2008), Dirección Estratégica, Buenos Aires, Argentina, Universidad Nacional de Quilmes.
- González, E. G. (2016). BIG DATA, PRIVACIDAD. Madrid: AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS.
- http://www.ujaen.es/investiga/tics_tfg/revi_sistematica.html
- https://azure.microsoft.com/es-es/services/data-factory/
- https://gallery.azure.ai/Solution/Campaign-Optimization-with-SQL-Server
- https://www.tynmagazine.com/banco-galicia-camino-a-ser-una-organizacion-data-driven/
- https://www2.deloitte.com/es/es/pages/risk/articles/covid-19-gestion-fraude-entidades-financieras.html
- Jimmy Janlen, agile coach en Spotify: https://www.youtube.com/watch?v=Tj-lavaMkxU
- Mayer, T., & Cyment , A. (2014). Por Un Scrum Popular: Notas para una Revolucion Agil. UK.
- Mayer, T., & Cyment , A. (2014). Por Un Scrum Popular: Notas para una Revolucion Agil. UK.
- Naya, S. (2018). Nuevo paradigma de big data en la era de la industria 4.0. Coruña: Disponible en: http://www.revistatog.com/num27/pdfs/editorial2.pdf.
- Puebla, J. G. (2018). Big Data y nuevas geografías: la huella digital. Madrid: Universidad Complutense de Madrid.
- © RapidMiner GmbH . (2020). rapidminer.com. Obtenido de https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html

- revistagq. (2020). Las 10 empresas tecnológicas más importantes de 2019 que han dominado nuestro ocio y consumo. revistagq. Obtenido de https://www.revistagq.com/noticias/articulo/empresas-tecnologicas-mas-importantes-2019
- Revuelta Bayod, M. (2018). Big Data: crisis y nuevos planteamientos en los flujos de comunicación. Revista de comunicación audiovisual.
- Revuelta Bayod, M. (2018). Big Data: crisis y nuevos planteamientos en los flujos de comunicación. Revista de comunicación audiovisual.
- S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, N. Kruschwitz (2011), Mit Sloan Management Review, USA
- Vijay, K., & Bala, D. (2015). Predictive Analytics and Data Mining. Waltham: Elsevier Inc.