INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE POSTGRADO

ENGAGEMENT PERFORMANCE EN CAMPAÑAS DE MARKETING

AUTOR/ES: Montero, Nicolás Ezequiel (Leg. Nº 503565)

DOCENTE/S TITULAR/ES O TUTOR/ES: Arjones, Gustavo

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIA DE DATOS

BUENOS AIRES

PRIMER CUATRIMESTRE, 2021

Trabajo Final Integrador

Introducción

El siguiente trabajo de investigación se aplica sobre el área de Marketing de una empresa que se dedica a la venta de servicios y productos informáticos.

Esta área se encarga de generar campañas de Marketing, utilizando diferentes canales, para poder captar la mayor cantidad de clientes que se interesen en adquirir los productos o servicios promocionados.

Los directivos de la empresa analizaron la performance de las diferentes campañas en los diferentes canales y detectaron que ciertas campañas no tenían el resultado esperado, con porcentajes muy bajos o nulos referentes al engagement de los clientes (midiéndose a través de dos métricas específicas: "Responses Creating Leads" y "Converted Leads From Lead Creating Responses").

Se busca analizar la performance actual de las campañas de marketing, entender las variables que entran en juego a la hora de obtener los resultados de engagement y abordando el problema a través del desarrollo de algoritmos de Machine Learning poder predecir si un cliente va a interesarse por los productos y servicios de la empresa. De esa manera se podrán obtener mejores resultados para las dos variables de engagement y por consecuencia generar más ganancias en la aplicación de cada campaña de Marketing. Esos algoritmos se van a validar y probar, verificando su nivel de efectividad para finalmente escoger aquel que sea el más certero.

Revisión Bibliográfica

La empresa de servicios informáticos no sólo provee servicios a sus clientes como acompañamiento en implementaciones y administración de productos, de migraciones a la nube como productos on premise y assesments frente a resolución de problemas y/o mejoras en performance, sino también venta de productos.

El área de marketing de la empresa se encarga de crear campañas a través de diferentes canales para poder acaparar la venta de servicios o productos a sus clientes, lo cual es llamado el engagement. Algunos de los diferentes canales que se encuentran habilitados al día de hoy para poder tener contacto con los clientes son:

- **Eventos:** Son eventos presenciales que se realizan en diferentes ciudades del mundo, donde los clientes tienen la posibilidad de poder ver sesiones donde se muestran los diferentes productos y servicios que la empresa brinda. Luego del evento los clientes pueden contactarse con el área de preventas para pedir una sesión personalizada, donde es factible la compra. El engagement se mide a partir de contabilizar cuantas personas se contactan con preventas y ventas mencionando que fue a partir de tal evento.
- **Mails:** Consiste en el envío de mails hacia un determinado grupo de clientes previamente segmentado. Cada email pertenece a una campaña diferente apuntando a la venta de servicios o productos específicos, existiendo en el email diferentes links a los que el cliente puede acceder,

donde posteriormente puede finalizar comprando el producto o el servicio que el mail ofrece. Luego se evalúa el engagement teniendo en cuenta la cantidad de personas que hicieron clic en el link del e-mail.

- **Advertising:** Consiste en diferentes métodos de propaganda de productos o servicios como por ejemplo revistas del área tecnológica donde se ofrecen estos mismos brindando un teléfono del área de ventas. El engagement se mide a partir de contabilizar cuantas personas se contactan con el área de ventas mencionando que fue a partir del advertising en por ejemplo una revista tecnológica.
- **Redes Sociales:** Consiste a través de las diferentes redes sociales que la empresa maneja (twitter, Facebook e Instagram) publicar contenido que sea atractivo para los clientes para que puedan poder requerir demos de productos o contratar servicios. El engagement se mide dependiendo de cada una de las redes sociales, ya que se miden las interacciones como las visitas, los clicks hechos sobre contenido, y también si a partir de ver ese contenido, si el cliente se comunica con el área de ventas.

Las dos métricas de performance que miden el engagement son:

- **Responses_Creating_Leads**: La métrica define si hubo una respuesta de un cliente a una campaña de marketing, buscando conocer más o adquirir un producto o servicio. Sus valores son 1 (cuando hubo esa respuesta) y 0 (cuando no hubo respuesta).
- **Leads_Converted_from_Leads_Creating_Responses**: La métrica define si una lead se convierte en una oportunidad (una compra o adquisición de un producto o servicio). Sus valores son 1 (cuando la lead se convierte en una oportunidad) y 0 (cuando no se convierte en una oportunidad).

A principio del tercer trimestre del año fiscal 2020 y tras la contratación de un nuevo líder en el cargo de Vice Presidente, a partir del 1 de Enero del año 2020, del área de Marketing, comenzaron a haber reuniones junto a los Senior Directores para evaluar cómo se están ejecutando las diferentes estrategias de Marketing en la empresa y cuáles fueron los resultados en los últimos dos años fiscales. A partir de dichas reuniones comenzó a analizarse si las campañas de marketing estaban generando el mínimo revenue esperado. El resultado fue que no, y que no se está teniendo el engagement esperado por parte de los clientes. Incluso se encontró que hay campañas que tienen engagement rate 0, con lo cual el nuevo Vice Presidente pidió identificar qué tipo de campañas son y en caso de que no actúen como se espera que directamente se las quite.

Comenzando un análisis preliminar de aquellas campañas que han tenido engagement 0 o casi nulo, se cree que las campañas que menos engagement generan son aquellas ejecutadas a través de emails dado que actualmente hay problemas de calidad de datos y actualización de datos de clientes. Si bien se generan listas de segmentación para poder clasificar que campañas se deben enviar a que clientes de

acuerdo a diferentes características que estos cumplen, como compras previamente hechas o links de la página de la empresa a la que ha accedido el cliente, ocurre que en la tabla de datos de los contactos, o bien no se tienen actualizados los emails de los mismos, o esos contactos no pertenecen a esas empresas. Como menciona McKinsey y Company (2013), "un estudio de más de 250 engagements por 5 años reveló que las compañías que ponen a los datos como prioridad en sus campañas de marketing, mejoran el retorno entre un 15-20%"

A este problema se suma que al no tener un seguimiento de los clientes más periódico hace que la empresa no mantenga actualizado que productos o servicios el cliente estaría interesado en acceder en la actualidad.

Como menciona Chintagunta, Hanssens, Hauser (2016) "La segmentación es un desglose de los perfiles de los consumidores en entidades homogéneas que comparten criterios comunes como el comportamiento de compra e historiales de transacciones".

Como menciona Abakouy, En-Naimi y El Haddadi (2017) "un gran porcentaje de Marketers termina seleccionando a sus contactos por intuición" más que por una selección exhaustiva y personalizada. Abakouy et al. (2017) explican que "para mejorar la performance de las campañas de email se requiere ir de un envío más masivo de emails, donde se atacan a una cantidad muy grande de clientes sin saber claramente si la información que se le envía al cliente es información que quiere o esperaría recibir, a un envío más específico y personalizado con mensajes más relevantes teniendo en cuenta la historia de compras de los clientes".

También mencionan que "el proceso de Emails Personalizados en Marketing es enviar la oferta correcta, en el momento correcto a la persona correcta basado en su profile. El objetivo principal del envío personalizado de emails es para identificar la necesidad del cliente y ofrecer productos y servicios que sean específicos para ese cliente en particular" Abakouy et al. (2017).

Un estudio de Marketing de Experian sugiere que "los emails personalizados incrementan el porcentaje de apertura en un 29% y el porcentaje de clicks en un 41%" Experian (2013).

El poder aplicar técnicas y algoritmos de machine learning ayudará a poder identificar a partir de comportamiento pasado de los clientes, cuál será su acción frente a diferentes campañas personalizadas que la empresa decida enviarles. Cómo mencionan Abakouy et al. (2017) "a través de un análisis del comportamiento actual se puede anticipar futuras acciones y por lo tanto poder ofrecerle al cliente ofertas híper adaptadas. El Marketing predictivo podría incluso responder a necesidades incluso no formuladas por los clientes. "

Como menciona eMarketer (2014) "El 81% de los clientes tienden a realizar compras cuando reciben un mail basado en compras previas", esperando así poder incrementar el engagement a partir de la aplicación de modelos predictivos.

Definición del problema

"En el área de Marketing de una empresa de servicios informáticos no todas las campañas realizadas

obtienen el engagement de clientes esperado y necesario.”

Justificación

Se necesita mejorar el engagement de los clientes, obteniendo un mejor rendimiento de las métricas “Responses Creating Leads” y “Converted Leads From Lead Creating Responses”. Pudiendo predecir en qué casos el cliente está interesado en un producto o servicio de la compañía para poder “atacar” de manera más eficaz a través de las campañas de Marketing, obteniendo como resultado final la compra o adquisición del producto o servicio.

Alcance de la Investigación

a- Los usuarios directos de esta investigación serán los directivos de la empresa que son quienes evalúan la performance de las campañas ejecutadas.

b- Los resultados serán utilizados por el área de Marketing y el/los algoritmos de machine learning para generar un modelo de predicción, serán brindados a compañeros que actualmente están trabajando en el área de Machine Learning dentro del área de Marketing.

Limitación

Los resultados obtenidos deberán ser aprobados por el director del área de marketing y el director del grupo de machine learning para poder ser utilizados dentro del equipo.

Hipótesis

“Las campañas de Marketing lanzadas para las regiones de Norte America, Latinoamerica y Europa, Medio Oriente y Africa que menos engagement generan en la empresa que brinda servicios informáticos son aquellas que se ejecutan a través de emails, obteniendo un bajo porcentaje de creación de leads y conversión de estas mismas a oportunidades ”

Objetivos

Desarrollar un modelo predictivo que contenga un algoritmo de Machine Learning que permita mejorar la performance de las métricas Responses Creating Leads y Converted Leads From Lead Creating Responses.

Objetivos específicos

- Desarrollar diferentes algoritmos de predicción.

- Validar y probar los diferentes algoritmos.
- Verificar el nivel de efectividad de los diferentes algoritmos.
- Determinar el algoritmo más efectivo y diseñar la interfaz del modelo.

Metodología

Para poder llevar a cabo la solución del problema planteado se utilizan técnicas de Machine Learning, de aprendizaje supervisado para poder entrenar un modelo que ayude a predecir si un cliente, de una compañía adquirirá un servicio o producto

La herramienta que se va a utilizar será Python para codificar la solución, utilizando dataframes para guardar la información utilizada para lograr la solución al problema, y librerías relacionadas a gráficos para poder mostrar los resultados obtenidos entre los diferentes modelos predictivos que se apliquen para el análisis y finales deducciones y conclusiones de los mismos.

Resultados

Los datos utilizados para el estudio del problema presentado contienen información de las compañías que fueron alcanzadas por las últimas campañas de marketing lanzadas por email, y que específicamente compran productos de la empresa que brinda servicios informáticos habiéndolas segmentado por:

- Región: sólo se analizaron compañías de las regiones de EMEA (Europa, Medio Oriente y África), LAD (Latinoamérica) y NA (Norteamérica);
- Industria: compañías de la industria de la minería, petróleo y gas;
- El tamaño de las compañías: caracterizado como MidSize, Lower Midsize y Above Midsize. Above Midsize son aquellas compañías que tienen un número de empleados mayor a 1500 y ganancias anuales mayores a $500 millones.
  Midsize son aquellas compañías que tienen un número de empleados entre 300 a 1500 y ganancias anuales de entre $100 millones a $500 millones.
  Lower Midsize son aquellas compañías que tienen un número de empleados menos a 300 y ganancias anuales menores a $100 millones.

Dentro de esa segmentación se tuvieron en cuenta variables que las definen a estas compañías como el país de origen, ganancias anuales, segmento del mercado al que pertenece la compañía, industria, número de empleados, si la compañía pertenece a una cuenta estratégica (aquellas compañías que han hecho mayores compras en los dos últimos años), como así también características respecto al contacto de esa compañía como a que departamento pertenece, el nivel dentro de la compañía y si ese contacto se puede contactar a partir de email o teléfono.

El dataset utilizado en este trabajo final integrador se ha restringido a no mostrar ningún dato

confidencial de ninguna compañía como así tampoco ningún dato de los contactos de estas mismas, teniendo como datos un total de 21489 registros y 26 atributos (variables) que a priori se consideraron importantes para el mismo y la creación de un modelo de predicción.

	COMPANY_KEY	COUNTRY	REGION	ANNUAL_SALES_USD	MARKET_SEGMENT	ONE_VOICE_INDUSTRY	ONE_VOICE_SEGMENT	COMPANY_REVENUE_BAND	MARKET_SEGMENT_BAND	NUMBER_OF_EMPLOYEES
0	58620980099	UNITED STATES	NaN	7.062920e+08	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	500-1B	500-1B	1900.0
1	91994225099	UNITED STATES	NaN	9.415000e+09	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	2B+	2B+	3177.0
2	45783157499	UNITED STATES	NaN	1.444477e+10	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	2B+	2B+	2800.0
3	45783157499	UNITED STATES	NaN	1.444477e+10	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	2B+	2B+	2800.0
4	45783157499	UNITED STATES	NaN	1.444477e+10	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	2B+	2B+	2800.0
...	...	...	...	...	...	...	...	...	...	...
21484	80544238499	UNITED STATES	NaN	1.024100e+10	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	2B+	2B+	26662.0
21485	33200299699	NIGERIA	EMEA	2.000000e+07	Lower Midsize	Mining, Oil and Gas	Support Activities for Mining	10-50M	10-50M	31.0
21486	80544238499	UNITED STATES	NaN	1.024100e+10	Above Midsize	Mining, Oil and Gas	Support Activities for Mining	2B+	2B+	26662.0
21487	50944087399	EGYPT	EMEA	2.300000e+10	Above Midsize	Mining, Oil and Gas	Oil and Gas Extraction	2B+	2B+	35000.0
21488	54971480699	BRAZIL	LAD	4.500000e+10	Above Midsize	Mining, Oil and Gas	Oil and Gas Extraction	2B+	2B+	68829.0

21489 rows × 26 columns

KEY_ACCOUNT_FLAG	EMAILABLE_FLAG	PHONABLE_FLAG	CONTACTABLE_FLAG	VANITY_DEPARTMENT	VANITY_LEVEL	VANITY_SPECIALITY	CHANNEL_NAME	LEAD_NUMBER	RESPONSE_ID
NaN	Y	Y	Y	Technology	Professional	Architecture	Others	AJN8B5	48520885.0
N	Y	Y	Y	Technology	Professional	General Technology	Others	AKWQJT	46379362.0
N	Y	Y	Y	Technology	Manager	Database Administration	Others	AFQ4F8	35138029.0
N	Y	Y	Y	Technology	Manager	Database Administration	Others	AJZYDC	35138029.0
N	Y	Y	Y	Technology	Manager	Database Administration	Others	AKXXWG	35138029.0
...	...	...	...	...	...	...	...	...	...
N	Y	Y	Y	Technology	Senior Professional	Systems	NaN	NaN	29785341.0
NaN	Y	Y	Y	NaN	Blank Title	NaN	NaN	NaN	30246191.0
N	Y	Y	Y	NaN	Blank Title	NaN	NaN	NaN	29769354.0
NaN	Y	Y	Y	Technology	Senior Professional	General Technology	NaN	NaN	35072004.0
N	Y	Y	Y	Technology	Professional	Analyst	NaN	NaN	56975904.0

RESPONSE_TYPE_ROLLUP_LEVEL_2	RESPONSE_TYPE_ROLLUP	RESPONSE_TYPE	Market_Segment_Band	Responses_Creating_Leads	Converted_Leads_From_Lead_Creating_Responses
Event	Online Events	iSeminar Webshow Attended	500	0	0
Form Submit	Oracle Promotions	Oracle Promotions	2000	0	0
Event	Session	Session Attended	2000	0	0
Event	Session	Session Attended	2000	0	0
Event	Session	Session Attended	2000	0	0
...	...	...	...	...	...
Form Submit	Whitepaper Downloaded	Whitepaper Downloaded	2000	0	0
Event	Attendees	Event Attended	10	0	0
Form Submit	Whitepaper Downloaded	Whitepaper Downloaded	2000	0	0
Event	Registered	Event Registered	2000	0	0
Event	Tradeshow Attendees	Event Tradeshow Engaged	2000	0	0

Analizando el dataset se decidió eliminar aquellas variables que se consideraron no relevantes para ayudar en la predicción, esas variables fueron:

'LEAD_NUMBER' -> debido a que es un número único por registro.
'RESPONSE_TYPE_ROLLUP_LEVEL_2' -> debido a que se tiene el campo RESPONSE_TYPE siendo mas granular.
'RESPONSE_TYPE_ROLLUP' -> debido a que se tiene el campo RESPONSE_TYPE siendo más granular.
'MARKET_SEGMENT_BAND' -> debido a que se tiene el campo Market Segment y se consideró no apropiado utilizar ambos.
'COMPANY_REVENUE_BAND' -> debido a que se tiene el campo Annual Sales USD y se consideró no apropiado utilizar ambos.
'RESPONSE_ID' -> debido a que es un número único por registro.
'VANITY_SPECIALITY' -> debido a que se tienen los campos VANITY DEPARTMENT y VANITY LEVEL y se consideró que con ese nivel de detalle sería suficiente.

Una vez que sólo quedaron aquellos campos (atributos) que se consideraron importantes para utilizar como variables para predecir las dos métricas (Responses_Creating_Leads y Leads_Converted_from_Leads_Creating_Responses), se procedió a evaluar si en el dataset existían valores nulos para las variables predictoras.

Se encontró que para varias de las variables existían valores NULL. El siguiente listado muestra la cantidad de registros NULL para cada una de las variables:

COMPANY_KEY	0
COUNTRY	0
REGION	14171
ANNUAL_SALES_USD	143
MARKET_SEGMENT	0
ONE_VOICE_INDUSTRY	0
ONE_VOICE_SEGMENT	0
NUMBER_OF_EMPLOYEES	12
KEY_ACCOUNT_FLAG	13567
EMAILABLE_FLAG	0
PHONABLE_FLAG	0
CONTACTABLE_FLAG	0
VANITY_DEPARTMENT	2474
VANITY_LEVEL	0
CHANNEL_NAME	5633
RESPONSE_TYPE	0
Market_Segment_Band	0
Responses_Creating_Leads	0
Converted_Leads_From_Lead_Creating_Responses	0

Se tomaron diferentes decisiones para cada una de las variables en cómo resolver los valores NULL:

Para las variables VANITY_DEPARTMENT y CHANNEL_NAME se decidió como valor default 'Unspecified' para aquellos casos que son NULL.

Para la variable ANNUAL_SALES_USD se decidió calcular la media y que ese fuese el valor default para ese campo.

Para la variable NUMBER_OF_EMPLOYEES, dado que sólo fueron 12 casos que se encontraban en NULL, se consideró como apropiado que el valor que lo reemplace fuera el mínimo dentro de los valores existentes para esa variable.

Para la variable KEY_ACCOUNT_FLAG, ya que es una variable que si la cuenta entra dentro de la categoría de KEY_ACCOUNT siempre se completa con el valor en Y, se decidió que el valor para los casos que son NULL sea 'N'.

Al evaluar la variable REGION se encontró que todos los casos pertenecían a los paises ESTADOS UNIDOS DE AMERICA y CANADA, por lo que el valor para esos casos NULL fue 'NA'.

A su vez, se encontró que las variables VANITY_DEPARTMENT y MARKET_SEGMENT tenían valores 'Unknown' por lo que se cambió su valor a 'Unspecified'.

Para poder aplicar el método de aprendizaje supervisado de clasificación Random Forest ("consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto. Cada árbol individual da una predicción de clase y la clase con más votos se convierte en la predicción del modelo" Yiu (2019)), se necesitó realizar un encoding de las variables categóricas. Para aquellas que eran flags, KEY_ACCOUNT_FLAG, EMAILABLE_FLAG, PHONABLE_FLAG y CONTACTABLE_FLAG se realizó un Ordinal Encoding, esto es, como sólo contiene 2 valores Y o N, el valor Y se transformó en 1 y el valor N en 0.

"Para aquellas variables categóricas donde no existe una relación ordinal, realizar un Ordinal Encoding

puede no ser suficiente, en el mejor de los casos, o engañoso para el modelo.

Si se realiza una relación ordinal a través de un Ordinal Encoding y se permite que el modelo asuma un orden natural entre categorías puede dar como resultado un rendimiento pobre en performance o resultados inesperados.” Brownlee (2020).

Para las variables del dataset, se puede aplicar One-Hot Encoding, donde se elimina la variable representada como entero y se agrega una nueva variable binaria para cada valor entero único en la variable.

Las variables que tuvieron que ser modificadas a través del One-Hot Encoding fueron:

REGION, MARKET_SEGMENT, ONE_VOICE_INDUSTRY, ONE_VOICE_SEGMENT, VANITY_DEPARTMENT, VANITY_LEVEL, CHANNEL_NAME, RESPONSE_TYPE.

Finalmente, el dataset modificado quedó con 119 variables. 117 variables predictoras y 2 variables a predecir.

	COMPANY_KEY	ANNUAL_SALES_USD	NUMBER_OF_EMPLOYEES	KEY_ACCOUNT_FLAG	EMAILABLE_FLAG	PHONABLE_FLAG	CONTACTABLE_FLAG	Market_Segment_Band	Responses_Creating_Leads
0	58620980099	7.062920e+08	1900.0	0	1	1	1	500	0
1	91994225099	9.415000e+09	3177.0	0	1	1	1	2000	0
2	45783157499	1.444477e+10	2800.0	0	1	1	1	2000	0
3	45783157499	1.444477e+10	2800.0	0	1	1	1	2000	0
4	45783157499	1.444477e+10	2800.0	0	1	1	1	2000	0
...	...	...	...	...	...	...	...	...	...
21484	80544238499	1.024100e+10	26662.0	0	1	1	1	2000	0
21485	33200299699	2.000000e+07	31.0	0	1	1	1	10	0
21486	80544238499	1.024100e+10	26662.0	0	1	1	1	2000	0
21487	50944087399	2.300000e+10	35000.0	0	1	1	1	2000	0
21488	54971480699	4.500000e+10	68829.0	0	1	1	1	2000	0

21489 rows × 119 columns

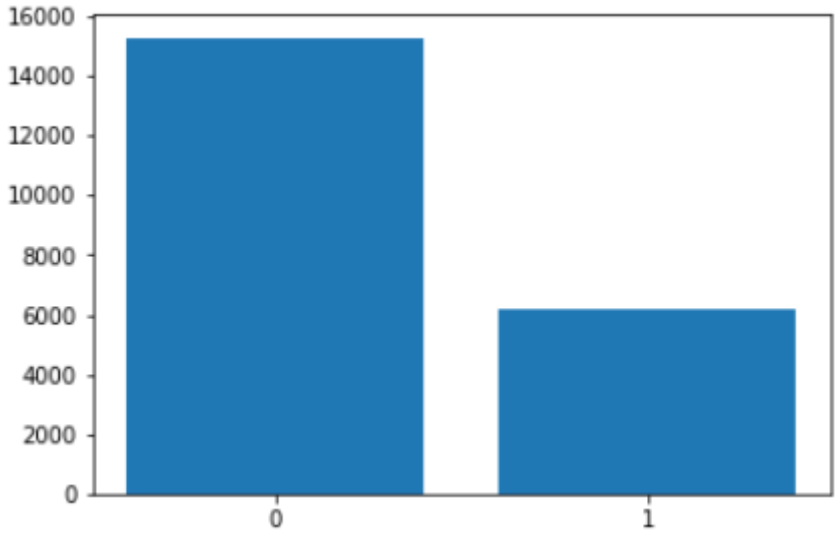
Se analizó la ocurrencia de las variables a predecir:

“Responses_Creating_Leads”

De un total de 21489 registros, 15297 tienen valor 0 y 6192 tienen valor 1.

```
0    15297
1     6192
```

Eso es un 71.19% de las ocurrencias tienen un valor 0 y el 18.81% tiene valor 1.



También se analizó la ocurrencia de los canales utilizados en las campañas agrupados por estas mismas para esta variable a predecir:

```
CHANNEL_NAME
Company websites - Tracked      1129
Email Tracked                   3392
External websites - Tracked      129
Others                         10418
Sales eVite                      638
Social Natural                   150
Unspecified                     5633
```

Como se observa del total, la mayor parte de los canales son “Others” y “Unspecified”. Pero de aquellos identificados las campañas por email es la que tiene más ocurrencias, siendo 3392 registros.

	CHANNEL_NAME	Avg_of_Responses_Creating_Leads
0	Company websites - Tracked	0.052539
1	Email Tracked	0.157848
2	External websites - Tracked	0.006003
3	Others	0.484806
4	Sales eVite	0.029690
5	Social Natural	0.006980
6	Unspecified	0.262134

Del total del dataset las campañas hechas por email son el 15,78% de ellas.

También se agruparon los registros del dataset por canal y por la métrica Responses Creating Leads:

CHANNEL_NAME	Responses_Creating_Leads	
Company websites - Tracked	0	611
	1	518
Email Tracked	0	2503
	1	889
External websites - Tracked	0	97
	1	32
Others	0	5865
	1	4553
Sales eVite	0	484
	1	154
Social Natural	0	104
	1	46
Unspecified	0	5633

Como se observa en la imagen se tiene para cada uno de los canales, cuantos registros pertenecen para el valor 0, no se creó una Lead, y cuantos pertenecen al valor 1, se logró crear una Lead.

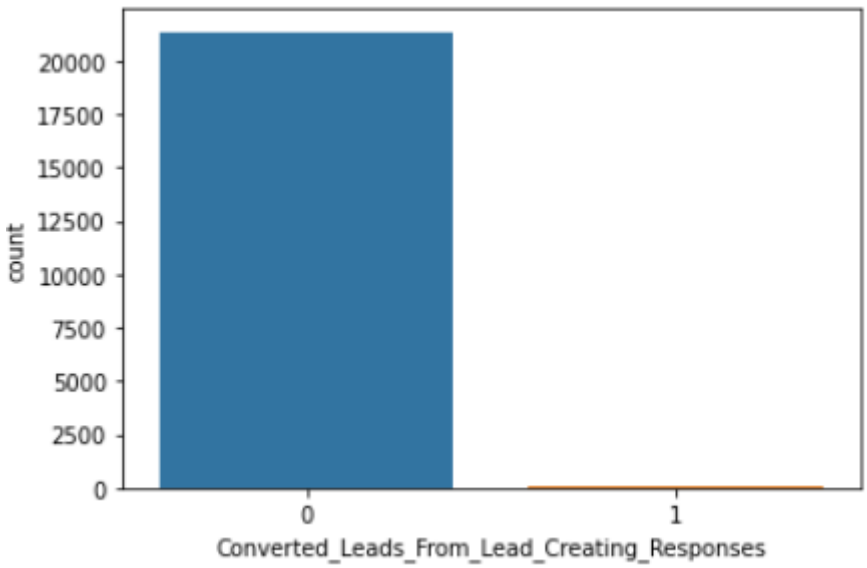
Analizando los valores para cada canal, se obtiene lo siguiente:

Para el canal “Company websites – Tracked”, el 45.88% se convierte en Lead, para “Others” el 43.70%, para “Social Natural” el 30.67%, para el canal “Email Tracked” el 26.21%, para “External websites – Tracked” el 24.81%, para “Sales eVite” el 24.14%.

“Leads_Converted_from_Leads_Creating_Responses”

De un total de 21489 registros, 21351 tienen valor 0 y 138 tienen valor 1.

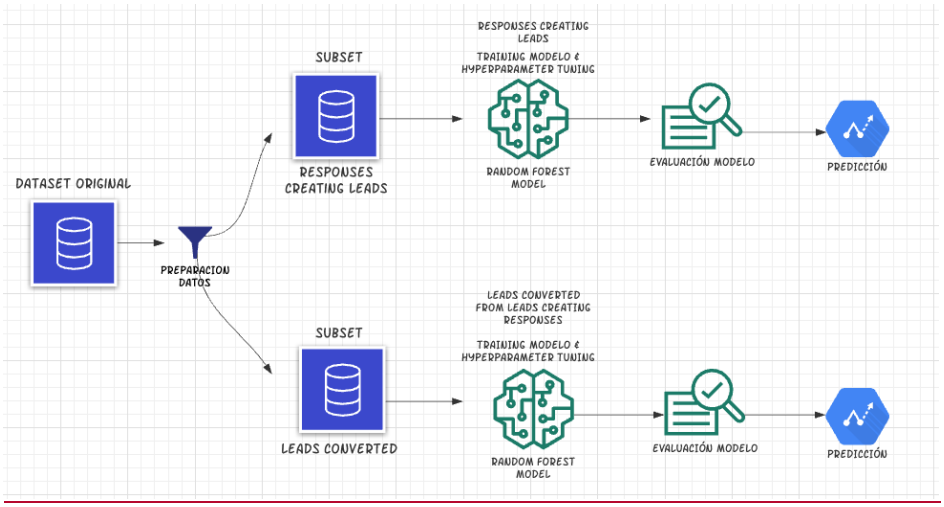
Eso es un 99.35% de las ocurrencias tienen un valor 0 y el 0.64% tiene valor 1.



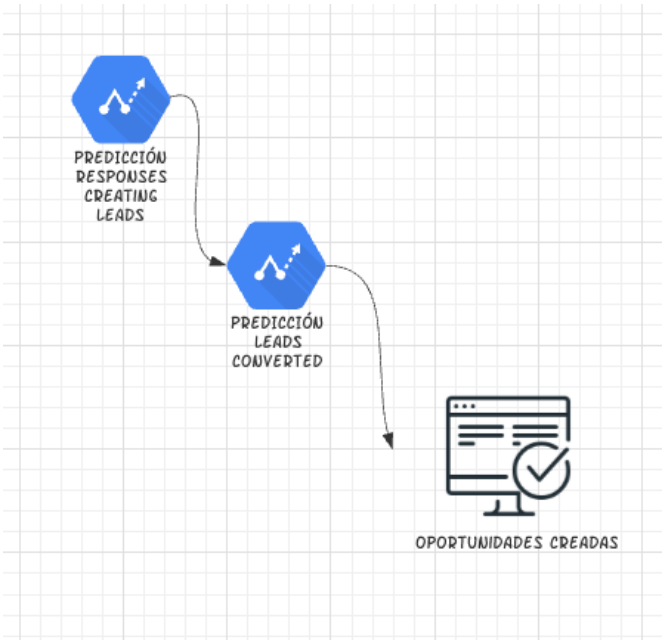
Debido a este análisis de las variables a predecir, por un lado “Responses_Creating_Leads” y por el otro “Leads_Converted_from_Leads_Creating_Responses” y dado que para que la segunda sea factible de que tenga un valor 1 (la lead se convirtió en una oportunidad) si o si la anterior variable tiene que ser 1, se decidió crear un segundo dataset a partir del ya obtenido luego de realizar todas las modificaciones anteriormente explicadas.

El segundo dataset tiene como requisito que la variable “Responses_Creating_Leads” sea 1. Por lo tanto, tiene una cantidad de registros de 6192.

A partir de la división de los dos datasets se siguió con el siguiente esquema para poder analizar y predecir las dos métricas a través del modelo de random forest:



Donde la predicción final de una se aplica a la predicción de la otra y finalmente a la obtención de los resultados de aquellas oportunidades creadas, como se muestra en el diagrama debajo:



A su vez, se eliminó del primer dataset la variable "Leads_Converted_from_Leads_Creating_Responses" ya que no iba a ser utilizada.

Para el caso del segundo dataset, también se analizó la ocurrencia de los canales utilizados en las campañas agrupados por estas mismas para esta variable a predecir:

CHANNEL_NAME	
Company websites - Tracked	518
Email Tracked	889
External websites - Tracked	32
Others	4553
Sales eVite	154
Social Natural	46

Como se observa del total, la mayor parte de los canales son "Others" y "Email Tracked", siendo para esta última, la cantidad de 889 registros.

	CHANNEL_NAME	Avg_of_Converted_Leads_From_Lead_Creating_Responses
0	Company websites - Tracked	0.083656
1	Email Tracked	0.143572
2	External websites - Tracked	0.005168
3	Others	0.735304
4	Sales eVite	0.024871
5	Social Natural	0.007429

Del total, el 14.35% de los registros son de campañas de “Email Tracked”.

Se agruparon los registros del dataset por canal y por la métrica Converted Leads from Lead Creating Responses:

CHANNEL_NAME	Converted_Leads_From_Lead_Creating_Responses	
Company websites - Tracked	0	511
	1	7
Email Tracked	0	847
	1	42
External websites - Tracked	0	32
Others	0	4469
	1	84
Sales eVite	0	153
	1	1
Social Natural	0	42
	1	4

Como se observa en la imagen se tiene para cada uno de los canales, cuantos registros pertenecen para el valor 0, la Lead no se convirtió en una Oportunidad, y cuantos pertenecen al valor 1, se logró convertir en una Oportunidad.

Analizando los valores para cada canal, se obtiene lo siguiente:

Para el canal “Social Natural”, el 8.7% se convierte en oportunidad, para “Email Tracked” el 4.72%, para “Others” el 1.84%”, para “Company websites – Tracked” el 1.35%, para “Sales eVite” el 0.65% y para “External websites – Tracked” el 0%.

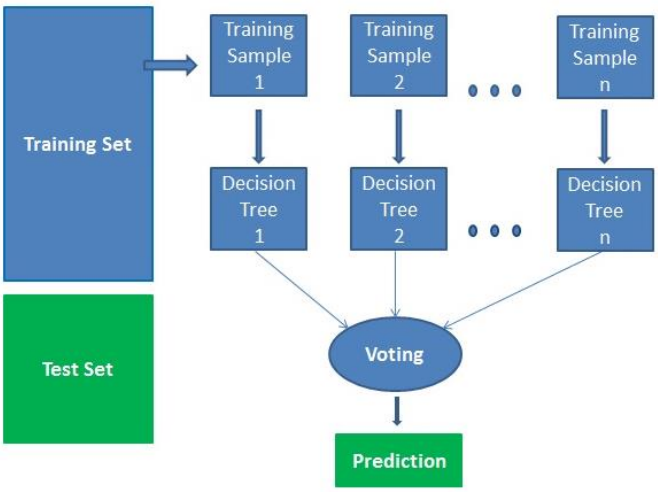
Como se pudo observar, en ninguna de las dos variables a predecir, las campañas de email fueron las que peor performance tuvieron para ambas métricas. En el caso de la métrica “Responses Creating Leads”, las campañas enviadas a través de email fueron las 4tas en conversión de Leads, y en el caso de la métrica “Converted Leads from Leads Creating Responses” es el segundo canal en las campañas que más convierten Leads en Oportunidades.

Como se mencionó anteriormente el método de clasificación utilizado fue Random Forest utilizando la librería Scikit-learn en Python. Así, primeramente se dividieron ambos datasets en features

(conteniendo todas las variables predictoras) y target(conteniendo la variable a predecir) para luego proceder a dividir ambos datasets en un conjunto de train y un conjunto de test, teniendo como tamaño del conjunto de testeo un 30% del mismo y 70% del conjunto de entrenamiento.

Como menciona Navlani (2018):
"El método de clasificación Random Forest trabaja de la siguiente manera:

1) Selecciona muestras random del dataset.
2) Construye un árbol de decisión para cada muestra y obtiene un resultado de predicción de cada árbol de decisión.
3) Vota para cada resultado de predicción.
4) Selecciona el resultado predicho con más votos como la predicción final."

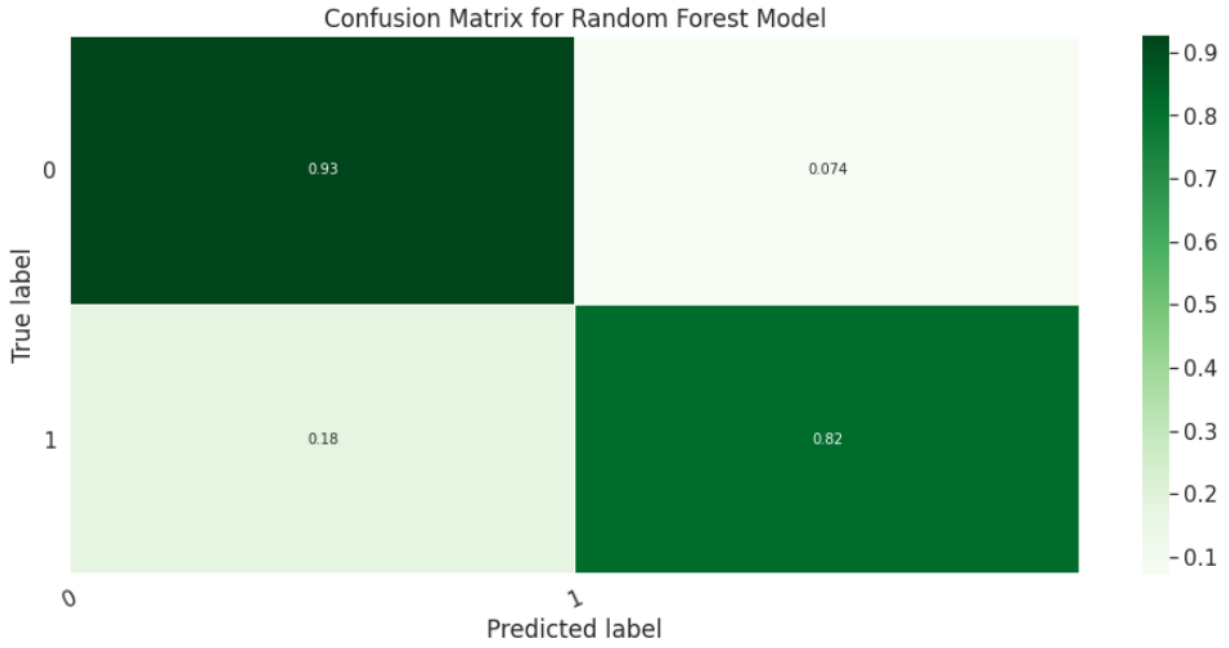


Predicción de la variable "Responses_Creating_Leads"

Modelo Base

Para la predicción de la primera variable se ejecutó el método utilizando como base los parámetros n_estimators (cantidad de árboles) = 10 y random_state (semilla) = 25.

La precisión de esta primera ejecución fue de un 89.65%. Realizando una matriz de confusión (herramienta para evaluar el desempeño del algoritmo de clasificación, dando una mejor idea de cómo está clasificando, a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación) se obtienen los siguientes valores:



Donde, analizando, el 93% de los casos que la variable era 0 lo acertó y el 82% de los casos donde la variable era 1 predijo que era 1.

También se evaluó un reporte de clasificación, donde se obtiene más detalle de la performance del algoritmo:

	precision	recall	f1-score	support
b'0'	0.93	0.93	0.93	4594
b'1'	0.82	0.82	0.82	1853
accuracy			0.90	6447
macro avg	0.87	0.87	0.87	6447
weighted avg	0.90	0.90	0.90	6447

Para entender un poco mejor lo que nos dice el reporte de clasificación se explica que significa cada una de sus métricas:

Precision es el número de valores correctamente identificados de una clase dividido por todas las veces que el modelo predijo esa clase. En el caso de nuestra variable, la puntuación de precisión sería el número de 1 correctamente identificados dividido por el número total de veces que el clasificador predijo 1, correcta o incorrectamente.

Recall es el número de valores de una clase que el clasificador identificó correctamente dividido por el número total de valores de esa clase. Para los 1, este sería el número de 1 reales que el clasificador identificó correctamente como tales.

La puntuación F1 es un poco menos intuitiva porque combina Precision y Recall en una sola métrica. Si Precision y Recall son altas, F1 también lo será. Si ambos son bajos, F1 será bajo. Si uno es alto y el otro

bajo, F1 será bajo. F1 es una forma rápida de saber si el clasificador es realmente bueno para identificar valores de una clase o si está encontrando atajos.

Entonces teniendo en cuenta el valor a predecir que nos importa que es el 1, tuvimos lo siguiente para nuestro algoritmo base:

Un accuracy del 89.65%, el 82% de las veces que predijo nuestro valor 1 lo acertó y las métricas Precision, Recall y F1 también en 82%.

Si planteamos que vamos a poder predecir el 82% de las veces cuando una response se convierta en Lead es un valor muy bueno.

Si bien la performance del algoritmo no fue mala, existe la posibilidad de ejecutar los hiperparametros que puedan ayudar a obtener una mejor performance incluso.

Para eso, se recurrió a dos tipos de hyperparameters:

Random Search Hyperparameters: donde de manera random se van combinando diferentes valores de los parámetros que el algoritmo de Random Forest puede tomar y devuelve el que obtiene mejor performance.

Grid Search with Cross Validation: una vez utilizado Random Search, el rango de valores a poder usar para los parametros se reduce y con Grid Search se obtienen dentro de esos valores reducidos para los parámetros, la mejor combinación.

Random Search Hyperparameters

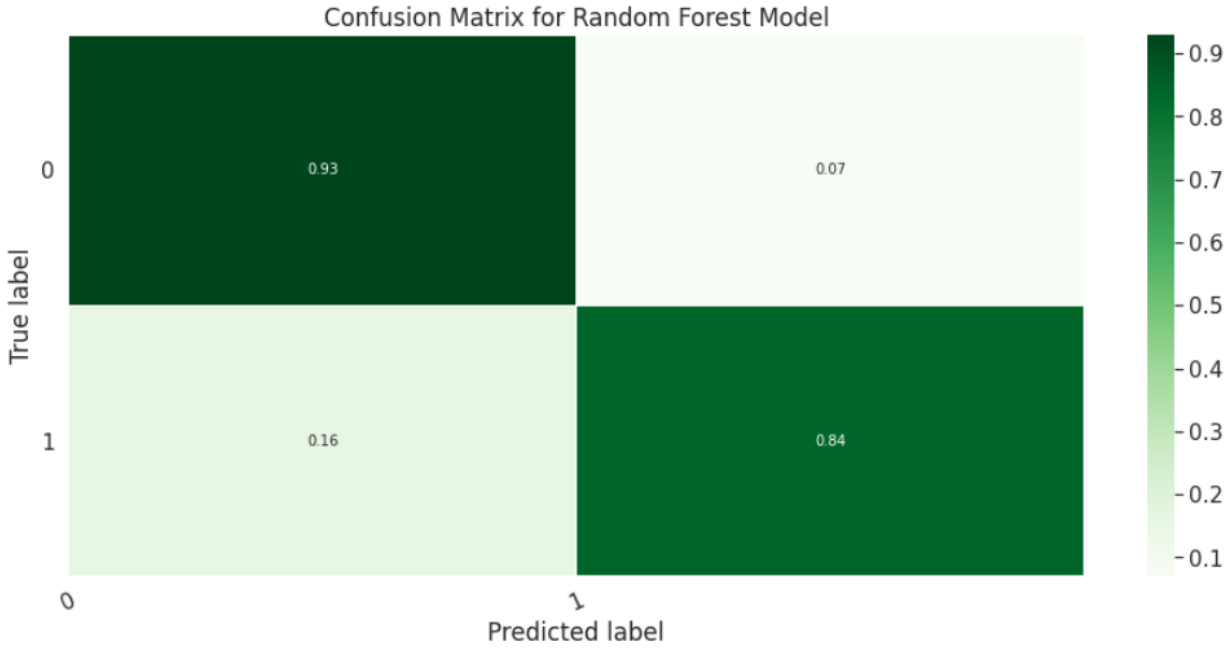
Utilizando Random Search, se obtuvo que los mejores valores para los parámetros eran los siguientes:

```
rf_random.best_params_
```

```
{'bootstrap': False,
 'max_depth': 70,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 1577}
```

Ejecutando el algoritmo de Random Forest con estos parámetros, más la semilla en 25, se obtuvieron los siguientes resultados:

La precisión de la segunda corrida del algoritmo fue de 90.45%. Si analizamos la matriz de confusión obtenida:



Se puede observar que para el valor que nos importa de nuestra variable a predecir, el 1, el 84% de las veces el algoritmo acertó que iba a ser un 1.

El reporte de clasificación:

	precision	recall	f1-score	support
b'0'	0.94	0.93	0.93	4594
b'1'	0.83	0.84	0.84	1853
accuracy			0.90	6447
macro avg	0.88	0.89	0.88	6447
weighted avg	0.90	0.90	0.90	6447

Donde se puede observar que para la métrica de Precision se obtiene un 83%, para Recall 84% y para F1-score 84% para el valor 1.

Nuestro nuevo algoritmo mejoró con respecto a la primera corrida con valores de los parametros base.

Grid Search with Cross Validation

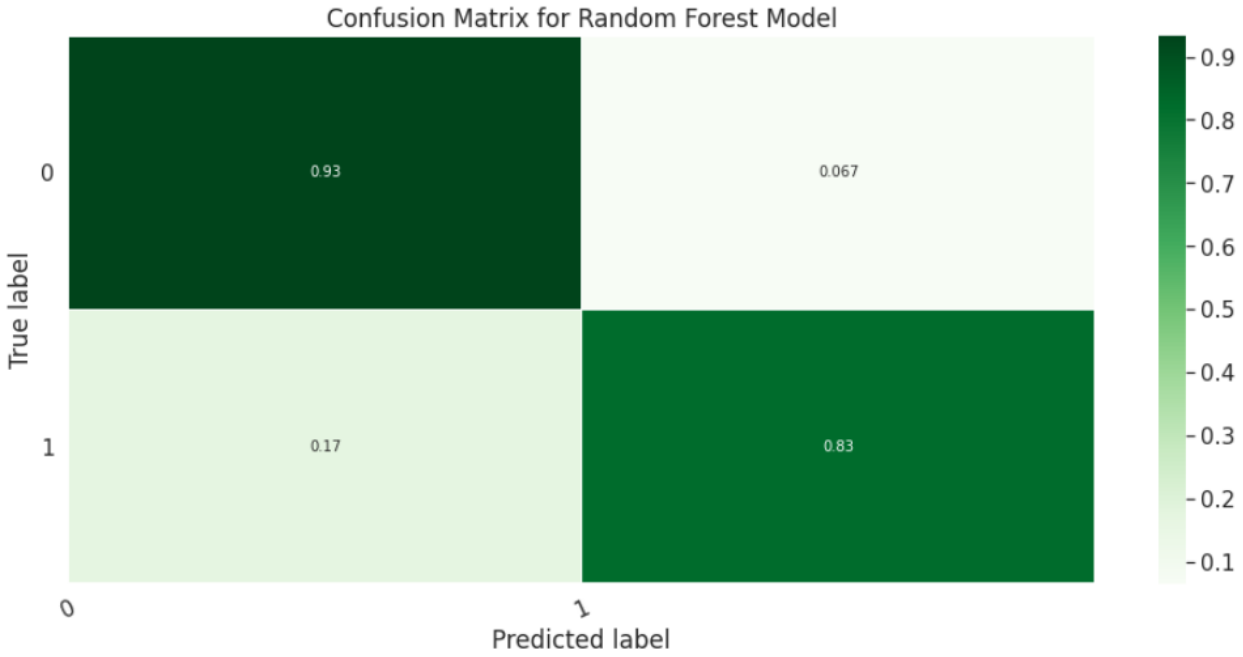
Para la corrida de Grid Search with Cross Validation se obtuvo que los valores optimos para los parámetros fueron:

```
grid_search.best_params_
```

```
{'bootstrap': False,
 'max_depth': 80,
 'max_features': 3,
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 2100}
```

Ejecutando el algoritmo de Random Forest con estos parámetros, más la semilla en 25, se obtuvieron los siguientes resultados:

La precisión de la tercera corrida del algoritmo fue de 90.20%. Si analizamos la matriz de confusión obtenida



Se puede observar que para el valor que nos importa de nuestra variable a predecir, el 1, el 83% de las veces el algoritmo acertó que iba a ser un 1.

El reporte de clasificación:

	precision	recall	f1-score	support
b'0'	0.93	0.93	0.93	4594
b'1'	0.83	0.83	0.83	1853
accuracy			0.90	6447
macro avg	0.88	0.88	0.88	6447
weighted avg	0.90	0.90	0.90	6447

Donde se puede observar que para la métrica de Precision se obtiene un 83%, para Recall 83% y para F1-score 83% para nuestro valor de 1.

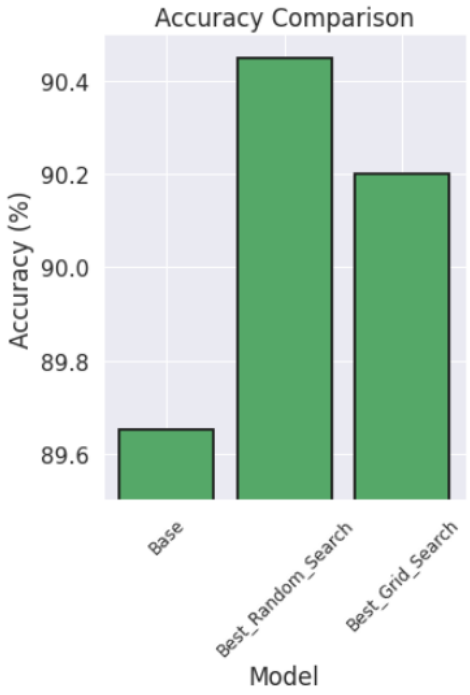
Nuestro nuevo algoritmo mejoró con respecto a la primera corrida con valores de los parametros base, pero no así con respecto a Random Search.

Comparación de modelos

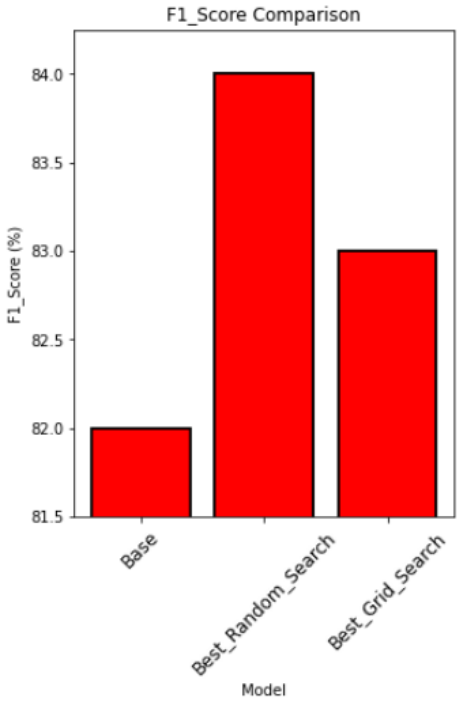
Si se comparan los 3 modelos ejecutados para la predicción de nuestra variable "Responses_Creating_Leads":

	Model	Accuracy	N_Arboles	Precision	Recall	F1_Score
0	Base	89.65	10	82	82	82
1	Best_Random_Search	90.45	1577	83	84	84
2	Best_Grid_Search	90.20	2100	83	83	83

Se puede observar que con respecto al modelo "Base", el accuracy del modelo Best_Random_Search mejora en un 0.8% y el accuracy del modelo "Best_Grid_Search" mejora en un 0.55%.



Con respecto a F1_Score, que combina a Precision y Recall, el modelo de "Best_Random_Search" obtiene el valor más alto con un 84%, 2 puntos más que el modelo Base (82%) y un punto más que el modelo "Best_Grid_Search" (83%).



Predicción de la variable “Converted_Leads_From_Lead_Creating_Responses”

Modelo Base

Para la predicción de la segunda variable se ejecutó el método utilizando como base los parámetros n_estimators (cantidad de árboles) = 10 y random_state (semilla) = 10.

La precisión de esta primera ejecución fue de un 98.12%. Realizando la matriz de confusión se obtienen los siguientes valores:

Donde, casi el 99,72% de los casos que la variable era 0 lo acertó y el 38% de los casos donde la variable era 1 predijo que era 1.

Con respecto al reporte de clasificación, donde se obtiene más detalle de la performance del algoritmo:

	precision	recall	f1-score	support
b'0'	0.98	1.00	0.99	1810
b'1'	0.78	0.38	0.51	48
accuracy			0.98	1858
macro avg	0.88	0.69	0.75	1858
weighted avg	0.98	0.98	0.98	1858

Para el valor 1, se obtuvieron los siguientes resultados para el algoritmo base:

Un accuracy del 98.12%, el 38% de las veces que predijo nuestro valor 1 lo acertó y las métricas Precision, Recall y F1 78%, 38% y 51% respectivamente.

En este caso, el 51% de las veces que una Lead se convierta en una Oportunidad vamos a poder predecirlo. Si bien un 51% no es un valor alto, para el negocio poder predecir en ese valor una oportunidad equivale a poder enfocar a los vendedores en esos clientes para concretar la compra o adquisición de un servicio.

Para la variable “Converted_Leads_From_Lead_Creating_Responses” se ejecutaron modelos con hiperparametros:

Random Search Hyperparameters

Utilizando Random Search, se obtuvo que los mejores valores para los parámetros eran los siguientes:

```
rf_random.best_params_
```

```
{'bootstrap': False,
 'max_depth': 100,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 733}
```

Ejecutando el algoritmo de Random Forest con estos parámetros, más la semilla en 10, se obtuvieron los siguientes resultados:

La precisión del algoritmo fue de 98.28%. Si analizamos la matriz de confusión obtenida:

Se observa que en el caso de los valores 0, el 99.89% de las veces acierta. Para el valor que más importa de nuestra variable a predecir, el 1, el 38% de las veces el algoritmo acertó que iba a ser un 1.

El reporte de clasificación:

	precision	recall	f1-score	support
b'0'	0.98	1.00	0.99	1810
b'1'	0.90	0.38	0.53	48
accuracy			0.98	1858
macro avg	0.94	0.69	0.76	1858
weighted avg	0.98	0.98	0.98	1858

Donde se puede observar que para la métrica de Precision se obtiene un 90%, para Recall 38% y para F1-

score 53% para el valor 1.

En este caso comparando contra el modelo base, el accuracy no mejoró, se mantuvo, pero si mejoraron los valores de precisión, recall y f1-score.

En este caso, el nuevo algoritmo mejoró con respecto al modelo base.

Grid Search with Cross Validation

Para la corrida de Grid Search with Cross Validation se obtuvo que los valores optimos para los parámetros fueron:

```
grid_search.best_params_
```

```
{'bootstrap': False,
 'max_depth': 80,
 'max_features': 2,
 'min_samples_leaf': 1,
 'min_samples_split': 3,
 'n_estimators': 1000}
```

Ejecutando el algoritmo de Random Forest con estos parámetros, más la semilla en 10, se obtuvieron los siguientes resultados:

La precisión de la tercera corrida del algoritmo fue de 98.33%. Si analizamos la matriz de confusión obtenida

Se puede observar que para el valor que nos importa de nuestra variable a predecir, el 1, el 40% de las

veces el algoritmo acertó que iba a ser un 1.

El reporte de clasificación:

	precision	recall	f1-score	support
b'0'	0.98	1.00	0.99	1810
b'1'	0.90	0.40	0.55	48
accuracy			0.98	1858
macro avg	0.94	0.70	0.77	1858
weighted avg	0.98	0.98	0.98	1858

Donde se puede observar que para la métrica de Precision se obtiene un 90%, para Recall 40% y para F1-score 55% para el valor 1 a predecir.

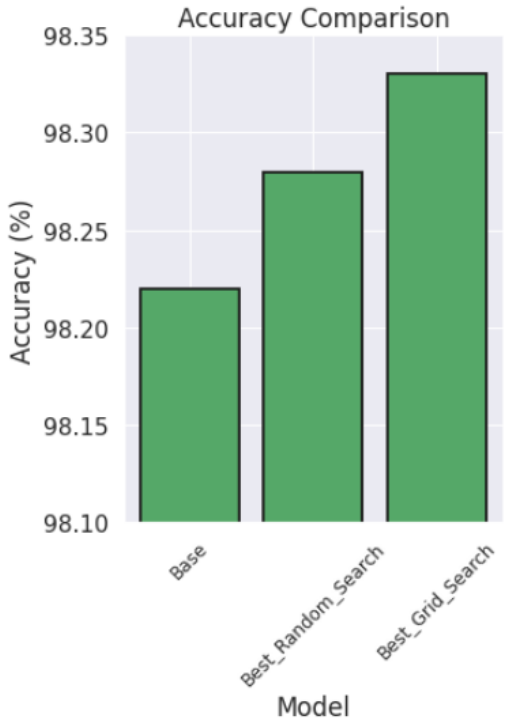
El nuevo algoritmo mejoró con respecto a las dos corridas previas (modelo base y modelo con parámetros de random search).

Comparación de modelos

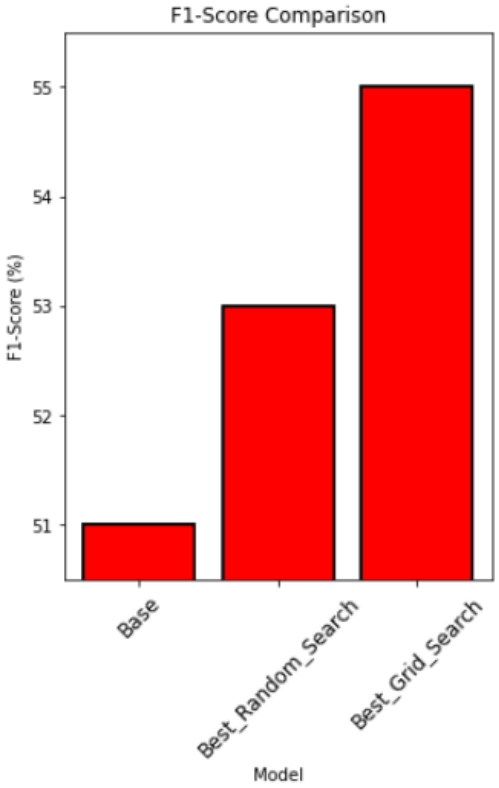
Si se comparan los 3 modelos ejecutados para la predicción de nuestra variable "Converted_Leads_From_Lead_Creating_Responses":

Model	Accuracy	N_Arboles	Precision	Recall	F1-Score
Base	98.22	10	78	38	51
Best_Random_Search	98.28	733	90	38	53
Best_Grid_Search	98.33	1000	90	40	55

Se puede observar que con respecto al modelo "Base", el accuracy del modelo Best_Random_Search mejora en un 0.06% y el accuracy del modelo "Best_Grid_Search" mejora en un 0.11%.



Con respecto a F1_Score, que combina a Precision y Recall, el modelo de “Best_Grid_Search” obtiene el valor más alto con un 55%, mientras que “Best_Random_Search” 53% y el modelo base 51%.



Conclusiones

El objetivo del presente trabajo final fue desarrollar un modelo predictivo que contenga un algoritmo de Machine Learning que permitiera mejorar la performance de las métricas Responses Creating Leads y Converted Leads From Lead Creating Responses para poder generar más engagement por parte de los clientes en las campañas lanzadas por la empresa de servicios y productos informáticos a través de los diferentes canales.

El análisis realizado permite entender cuales fueron los canales que mejor performance tuvieron en el último tiempo, dándole la posibilidad al equipo de Marketing de poder decidir para futuras campañas de enfocarse en uno o varios canales o de prescindir de uno de ellos dado que no se logran porcentajes altos de conversiones, como por ejemplo en el caso de la variable "Converted Leads from Lead Creating Responses" donde el canal "External websites – Tracked" obtuvo el 0% de conversiones a oportunidades. El canal de "Email Tracked" que fue el que se pensaba que obtenia peor performance de todos, en el caso de la métrica "Responses Creating Leads" era el 4to canal en obtención de conversión y en el caso de la métrica "Converted Leads from Lead Creating Responses" fue el 2do canal con mejor conversión.

Con respecto al estudio de la métrica "Responses Creating Leads" se evaluaron 3 modelos, Random Forest, Random Forest con Random Search y Grid Search with Cross Validation, de los cuales se obtuvo que el que mejor performance obtuvo en predecir si una respuesta del cliente se convertirá en Leads o no, fue el algoritmo Random Forest con Random Search Hyperparameters, donde se logró un accuracy del 90.45% y que en el 84% de las veces poder acertar que una Response se convertirá en una Lead.

Con respecto a la métrica "Converted Leads from Lead Creating Responses", también se evaluaron 3 modelos, donde el que mejor performance obtuvo fue "Grid Search with Cross Validation", obteniendo un 98.33% de accuracy, y pudiendo en el 55% de las veces poder acertar cuando una Lead se convertirá en Oportunidad.

Los valores obtenidos para ambas métricas, 90.45% y 84% para accuracy y f1-score para "Responses Creating Leads", y 98.33% y 55% en el caso de la métrica "Converted Leads from Lead Creating Responses" son valores muy buenos para el caso puntual del negocio dado que aportan la posibilidad de mejorar el rendimiento de las campañas y para los vendedores el modelo ayudará a poder identificar apropiadamente las Leads que deberán atender prioritariamente, como a su vez entender que características de las Leads son las que ayudan a convertir con mayor efectividad. Esto permitirá al área a que el sistema aprenda a que en las próximas campañas un tipo de Leads específicas van a ser aquellas en las que se debe enfocar el área de ventas primeramente sabiendo que es factible en un 84% para la métrica "Responses Creating Leads" y en un 55% para la métrica "Converted Leads from Lead Creating Responses" conviertan respectivamente en Leads y en Oportunidades con modelos con un accuracy del 90.45% y un 98.33% respectivamente.

Esta investigación es un primer paso para poder mejorar la performance de las campañas de Marketing dentro de la empresa, dejando abierta la posibilidad a seguir analizando otras posibles variables que jueguen un rol importante en obtener una mejor conversión de Leads y Oportunidades, cómo por

ejemplo sumar la información de quien es el vendedor para cada caso, para así poder analizar su performance histórica, entender si ha logrado los objetivos y poder por ejemplo atacar esta variable mejorando sus soft skills.

Referencias Bibliográficas, Artículos Científicos, Congresos o Revistas

- Redouan Abakouy, El Mokhtar En-Naimi, Anass El Haddadi, Noviembre 2017, Classification and Prediction Based Data Mining algorithms to Predict Email Marketing Campaigns.
- Redouan Abakouy, El Mokhtar En-Naimi, Anass El Haddadi, Elaachak Lotfi, Octubre 2019, Data-driven marketing: how machine learning will improve decision-making for marketers.
- Experian, 2013, How today's email marketers are connecting, engaging and inspiring their customers.
- McKinsey & Company, 2013, Smart analytics can tap up to 20% of lost ROI.
- eMarketer, 2014, Personalization Sees Payoffs in Marketing Emails.
- P. Chintagunta, D. M. Hanssens, J. R. Hauser, 2016, Marketing Science and Big Data, MARKETING SCIENCE Vol. 35, No. 3, pp. 1–2 ISSN 0732-2399 (print) I ISSN 1526-548X (online)
- T. Yiu, 2019, Understanding Random Forest. Recuperado de https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- J. Brownlee, 2020, Ordinal and One-Hot Encodings for Categorical Data. Recuperado de https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/
- A. Navlani, 2018, Understanding Random Forests Classifiers in Python. Recuperado de https://www.datacamp.com/community/tutorials/random-forests-classifier-python