

**UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA**

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN



**“SEGMENTACIÓN DE USUARIOS QUE VISITAN EL SITIO WEB DE
UNA EMPRESA UTILIZANDO LA REGRESIÓN LOGÍSTICA CON LA
TÉCNICA DE SOBREMUESTREO”**

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL

TÍTULO DE

INGENIERO ESTADÍSTICO E INFORMÁTICO

CARLOS MARCIAL TASAYCO SILVA

LIMA – PERÚ

2021

**La UNALM es titular de los derechos patrimoniales de la presente
investigación (Art. 24- Reglamento de Propiedad Intelectual)**

**UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA**

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

**“SEGMENTACIÓN DE USUARIOS QUE VISITAN EL SITIO WEB DE
UNA EMPRESA UTILIZANDO LA REGRESIÓN LOGÍSTICA CON LA
TÉCNICA DE SOBREMUESTREO”**

PRESENTADO POR

CARLOS MARCIAL TASAYCO SILVA

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR POR EL
TÍTULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO**

SUSTENTADO Y APROBADO ANTE EL SIGUIENTE JURADO

.....
MS. Víctor Manuel Maehara Oyata
PRESIDENTE

.....
Dr. Jaime Carlos Porras Cerrón
MIEMBRO

.....
Dr. César Higinio Menacho Chiok
MIEMBRO

.....
Dr. Jorge Chue Gallardo
ASESOR

LIMA – PERÚ

2021

DEDICATORIA

Dedico este trabajo de monografía a mi familia, por su incondicionable apoyo para mi persona, a pesar del espacio que nos separa y el tiempo de incertidumbre en el que vivimos actualmente.

AGRADECIMIENTO

En primer lugar, quiero agradecer a la empresa "Attachmedia" donde labore por un poco mas de 2 años, donde juntos encaminamos con éxito una gran cantidad de proyectos. Sin duda fue una experiencia que me enriqueció como profesional porque forme y lidere equipos con una gran variedad de habilidades, además me hizo mejor persona al estar en contacto con personas con grandes valores.

También quiero agradecer al profesor Jorge Chue por su gran compromiso en asesorarme en crear este trabajo de monografía sin duda este trabajo no seria ni la mitad de bueno sino fuera por su apoyo.

Por último, quiero agradecer a la universidad por permitirme formarme como profesional lo cual me ha ayudado a realizar todos los logros tanto profesionales como personales en mi vida adulta. Sino hubiera pasado por esta casa de estudios no hubiera tenido la oportunidad de conseguir empleo y estudiar en el extranjero, y lo mas importante jamás hubiera conocido a las personas con las que estoy formando una vida el día de hoy.

ÍNDICE GENERAL

I. INTRODUCCIÓN	9
1.1 Problemática	9
1.2 Objetivos	11
II. MARCO TEÓRICO	12
2.1. Segmentación	12
2.2. Regresión Logística.....	14
2.3. Técnica de Sobremuestreo	17
2.4. AUC	18
III. MARCO METODOLÓGICO.....	24
3.1. Alcance.....	24
3.2. Tipo de investigación	24
3.3. Población.....	24
3.4. Muestra.....	25
3.5. Recolección y modelamiento de datos	25
3.6. Variables	25
3.6.1. Variable Dependiente.....	25
3.6.2. Variables Independientes	26
3.7. Análisis.....	27
IV. RESULTADOS Y DISCUSIÓN	29
4.1. Resultados	29
4.1.1. Cargar las Librerías	29
4.1.2. Análisis univariante.....	30
4.1.3. Categorización de las variable cualitativas	33
4.1.4. Técnica de sobremuestreo	34
4.1.5. Modelo de regresión logística	35
4.1.6. Validación del modelo	39
4.1.7. Segmentación	40
4.2. Discusión.....	43
V. CONCLUSIONES	47
VI. RECOMENDACIONES.....	49

ÍNDICE DE FIGURAS

Figura 1. Capas de análisis de segmentación. Fuente: (Dolnicar, 2018)	13
Figura 2. Conjuntos de datos original y de datos después de aplicar SMOTE	18
Figura 3 Curva AUC. Fuente: (Jiménez, 2012)	20
Figura 4 Ejemplo de Web scrapping de una web	21
Figura 5 Arquitectura del almacenamiento de datos en la nube donde está Big Query	23
Figura 6. Código de Python para cargar las librerías Pandas, Numpy, Sklearn y Matplotlib	30
Figura 7. Cantidades de usuarios que realizan transacciones (1) y que no realizan transacciones (0)	30
Figura 8. Análisis univariante de la variable “canal”	31
Figura 9. Análisis univariante de la variable “tipo de dispositivo”	31
Figura 10. Análisis univariante de la variable “tipo de sistema operativo”	32
Figura 11. Análisis univariante de la variable “Usuario Nuevo”	32
Figura 12. Análisis univariante de la variable “Pagina de destino”	33
Figura 13. Código en lenguaje de programación Python para categorizar las variables cualitativas	33
Figura 14. Data con las nuevas variables binarias creadas	34
Figure 15 Distribución de usuarios por nro de visitas	34
Figura 16. Código en lenguaje de programación Python dónde se aplicará la técnica de sobremuestreo “Smote”	35
Figura 17. Clases balanceadas después de aplicar la técnica de sobremuestreo	35
Figura 18. Código en lenguaje de programación Python para realizar un modelo de regresión logística a la data de entrenamiento	36
Figura 19. Resultados del modelo de regresión logística con todas las variables de la data de entrenamiento	36
Figura 20. Resultados del modelo de regresión logística solo con las variables que dan información significativa al modelo original	37
Figure 21 Código en lenguaje de programación Python para obtener las probabilidades estimadas y obtener el área bajo la curva	37
Figura 22. Código en lenguaje de programación Python para obtener la matriz de confusión	39
Figura 23. Código en lenguaje de programación Python para realizar la curva ROC	40

Figura 24. Curva ROC	40
Figura 25. Diagrama de cajas y bigotes	41
Figura 26. Código en lenguaje de programación Python para obtener los percentiles 50 y 75	41

ÍNDICE DE TABLAS

Tabla 1. Comparación entre Recorrido de cliente y un modelo de toma de decisiones_	14
Tabla 2. Comparación entre la regresión logística y las redes neuronales_____	16
Tabla 3. Descripción de todas las variables del modelo de regresión logística_____	38
Tabla 4. Matriz de confusión_____	39
Tabla 5. Segmentos e intervalo de probabilidades_____	42
Tabla 6 Ejemplo de tres observaciones a las que se les aplico el modelo estadístico_____	42

RESUMEN

El trabajo de monografía se desarrolló en una empresa consultora líder en Latinoamérica especializada en soluciones analíticas para las áreas de marketing digital de empresas multinacionales. El trabajo consistió en la implementación y automatización de un modelo de regresión logística binaria para determinar los segmentos de los usuarios que visitan la Web de la empresa. Para realizar el modelo estadístico mencionado se empezó desde el análisis univariante de cada variable independiente, seguido por una técnica de sobremuestreo “SMOTE” para evitar el desbalance de las clases en la variable dependiente, se realizó además una matriz de confusión en la cual se obtuvo una precisión del 76%, hasta finalmente validar la predictibilidad analizando la curva ROC. Los resultados de la investigación ayudaron a demostrar y concluir que existen variables que son significativas para determinar si un usuario que visita la Web realiza una transacción. Por ejemplo: Los usuarios que usan canales orgánicos sin medios publicitarios como referencia y directo o pagados como las redes sociales contribuyen negativamente a la probabilidad de que el usuario haga la transacción, así como también se observó que tanto el tiempo de la visita o si el usuario visita la Web recurrentemente contribuyen a que la probabilidad de hacer la transacción se incremente. Finalmente, la segmentación que se realizó basado en las puntuaciones calculadas por la regresión logística binaria para tener tres segmentos bien diferenciadas que son alto, medio y bajo probabilidad. Al final del trabajo, la empresa aceptó y mostró su satisfacción con los resultados obtenidos.

Palabras claves: Regresión logística, SMOTE, univariante, segmentación, ROC

ABSTRACT

The monograph work was developed in a leading consulting company in Latin America specialized in analytical solutions for the digital marketing areas of multinational companies. The work consisted in the implementation and automation of a binomial logistic regression model to describe the segments of users who visit one of the client's website. Therefore, to perform the aforementioned statistical model, it started from the univariate analysis of each independent variable, followed by a "SMOTE" oversampling technique to avoid the imbalance of classes in the dependent variable, then with a matrix confusion with an accurate of 76% until finally validating the predictability by analyzing the ROC curve. The results of the research helped to demonstrate and conclude that there are variables that are significant in determining whether a user visiting the Web makes a transaction. Users who use organic channels without advertising media as a reference and direct or paid such as social networks negatively contribute to the likelihood of the user making the transaction. It was also observed that both the time of the visit or if the user visits the Web recurrently contribute to the probability of making the transaction is increased. Finally, the segmentation was performed based on the scores calculated by the binomial logistic regression to have three well-differentiated segments that are high, medium and low probability. At the end of the work, the company accepted and showed its satisfaction with the results obtained.

Keywords: Logistic Regression, SMOTE, univariate, segmentation, ROC.

I. INTRODUCCIÓN

La empresa de consultoría (EC) donde se desarrolló el trabajo de suficiencia profesional tiene más de 15 años de experiencia en el rubro ofreciendo los servicios de optimización en motores de búsqueda, publicidad digital, analítica digital, experiencia de usuario y proyectos de transformación digital. En el 2017, se creó el laboratorio de innovación y desarrollo para ofrecer soluciones informáticas para MYPES. Desde el 2020, la empresa inauguró una nueva oficina en Miami, aumentando su presencia hasta tres países: Perú, México y USA.

En la actualidad, la empresa tiene alrededor de 50 colaboradores y más de 20 clientes, además de realizar conferencias anuales con invitados internacionales reconocidos a nivel mundial, siendo estas últimas importantes centros de negocios para la captación de potenciales clientes. En este trabajo monográfico se denominará “cliente” a la empresa a quien la EC brinda sus servicios.

Desde el 2017, el cliente ha aumentado constantemente la inversión en medios digitales llegando a prácticamente a duplicarla en el año 2019 respecto al año anterior. Sin embargo, este incremento no se ve reflejado en las ventas llegando apenas a tener un incremento anual del 30% en el 2018 y apenas un 10% en el año 2019. Otro indicador que preocupa al cliente es la proporción de transacciones que terminaron en una venta sobre el número total de visitas a la página Web del cliente, este ha disminuido de 35% en el 2017 a aproximadamente 13% en el 2019. Estos resultados indican claramente que existe un problema de estrategia a la hora de segmentar las audiencias a los cuales se les envía la publicidad. Finalmente, por estudios de opinión que se han realizado se sabe que existe un problema de experiencia de usabilidad en el sitio Web ya que los usuarios entrevistados así lo indicaron.

1.1 Problemática

El problema identificado en la empresa de consultoría es la falta de segmentación de los usuarios de los clientes a quienes brinda sus servicios ocasionando que no se cumplan los

objetivos de ventas a pesar de que el nivel de inversión es cada vez mayor. La solución escogida es la aplicación de un modelo de regresión logística para realizar la segmentación y está justificado porque se puede automatizar fácilmente, tiene una sustentación científica que es altamente valorado por el cliente y entrega una solución bastante aceptable. El problema identificado por la EC es la falta de segmentación de los usuarios de los clientes a quienes brinda sus servicios ocasionando que no se cumplan los objetivos de ventas a pesar de que el nivel de inversión es cada vez mayor.

La alternativa de solución que fue considerada inicialmente fue realizar un trabajo de diseño de la página Web del cliente con la hipótesis que la misma no era adecuada para el tipo de usuario que ingresaba a visitarla. Esto ocasionó un cambio de diseño en la tienda que era de interés mejorar, pero las ventas no se incrementaron significativamente. Otra alternativa propuesta fue realizar un análisis de agrupación (K-medias) para segmentar a los usuarios en diferentes grupos y entender como era el comportamiento de estos para crear una publicidad más personalizada. Sin embargo, los grupos identificados eran muy heterogéneos con un grupo muy grande donde estaban casi el 70% de los usuarios y los demás grupos eran muchos más pequeños.

La solución elegida fue la automatización de un modelo de regresión logística que permita segmentar cada visita y saber que tipo de diseño Web es el apropiado para mostrarle al usuario. La razón de la elección de esta solución es porque se busca una forma de personalizar la experiencia de cada usuario. Por ejemplo: si la visita del usuario ingresa a un segmento de usuarios con alta probabilidad de compra se le ofrecerá una Web comercial mucho más agresiva para que la transacción sea exitosa; por el contrario, si ingresa a un segmento de usuarios con baja probabilidad de compra se le mostrara una Web informativa con un formulario de compra mucho menos complejo para que sea contactado por un asesor de

ventas por teléfono. Otras razones de esta elección son: el bajo costo de desarrollo, los resultados se pueden monitorear mediante tableros de control y la integración del modelo dentro del ambiente de desarrollo del cliente que era muy importante para ellos.

Con la solución propuesta se busca conseguir que la proporción de transacciones sobre visitas se incremente de 13% a por lo menos 25% y que la experiencia de los usuarios mejore debido a que se les ofrecerá un diseño personalizado.

El alcance de esta investigación tiene un espacio temporal de 30 días porque las estrategias se definen cada inicio de mes y con la data mencionada aquí. Esto se debe a que los resultados de un modelo de regresión logística binaria varia dependiendo de la data y posiblemente pueda cambiar las conclusiones si se volviera a repetir el experimento, lo cual se espera porque el objetivo es mejorar la tendencia en ventas que tenemos actualmente.

1.2 Objetivos

Objetivo general: Realizar la segmentación de los usuarios que llegan a la Web para ofrecerles una experiencia personalizada y de esta manera obtener más transacciones, basado en la probabilidad de que cada usuario haga una transacción.

Objetivos específicos:

- Categorizar a los usuarios en tres grandes grupos: alta probabilidad, media probabilidad y baja probabilidad de realizar una transacción, aplicando la regresión logística binaria y la tecnica de sobremuestreo SMOTE.
- Demostrar que existe distintos tipos de usuarios basados en la probabilidad de que hagan una transacción con sustento estadístico.
- Identificar las variables que determinan que un usuario haga una transacción o no.

II. REVISIÓN DE LITERATURA

2.1. Segmentación

En un estudio de Goyat se menciona que, las estrategias de marketing o campañas de marketing efectivas a menudo consisten en una combinación de varias tácticas de marketing que funcionan juntas de manera sinérgica para establecer su marca, reducir la resistencia de ventas y crear interés y deseo por su producto o servicio. Hoy en día, el marketing está en todas partes, formal o informalmente, las personas y las organizaciones participan en una gran cantidad de actividades que se denomina marketing. Todas las empresas quieren centrarse en los clientes dentro de su capacidad y con la intimidad de los clientes. Pues este mercado consiste en dividirse en grupos de consumidores o segmentos con necesidades y deseos distintos. Esta estrategia de dividir el mercado en grupos homogéneos se conoce como segmentación (Goyat, 2011).

Grishikashvili sostiene que en la era de la nueva economía digital la demanda de comprender cómo administrar y analizar la información comercial a gran escala (Big Data) de una manera eficaz y eficiente para el éxito empresarial es muy alta. La transformación de datos en información y conocimiento ayuda a configurar estrategias efectivas para gestionar el conocimiento para la investigación de mercado (Grishikashvili, 2014). Sin embargo, la segmentación todavía tiene un papel estratégico y táctico, siempre ocurre esto cuando se trata de implementar una segmentación de mercado:

- El papel estratégico de la segmentación sigue siendo importante.
- La implementación es tan dolorosa como siempre.
- Incluso los segmentos más inteligentes pueden ser difíciles de integrarse.

¿Por qué segmentar? Porque permite identificar segmentos donde los competidores ven un mercado masivo indiferenciado crea varias oportunidades para nuevas estrategias de marketing basadas en un mejor conocimiento de las necesidades y preferencias específicas de los clientes (Fonseca, 2011).

Finalmente, Dolnicar señala que las capas del análisis de segmentación del mercado son: extraer los segmentos de mercados inteligentemente, analizar la data y hacer que estos segmentos funcionen correctamente (Dolnicar, 2018).

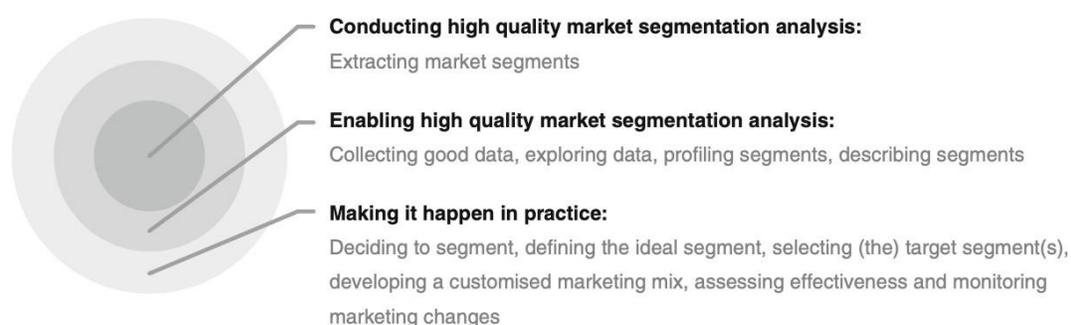


Figura 1. Capas de análisis de segmentación. Fuente: (Dolnicar, 2018)

Según Wolny, la segmentación de clientes puede generar múltiples recorridos de comprador y posiblemente diferentes experiencias de cliente, el recorrido del cliente se puede definir como "Una descripción de la experiencia del cliente donde diferentes puntos de contacto caracterizan la interacción de los clientes con una marca, producto o servicio de interés". La clasificación de interacciones a menudo no sigue una estructura lineal. También involucra una serie de canales y refleja las respuestas emocionales, conductuales y cognitivas presentes en el proceso. La tabla 1 compara los diferentes aspectos que distinguen los recorridos del cliente de los modelos de toma de decisiones tradicionales (Wolny, 2014).

Tabla 1. Comparación entre Recorrido de cliente y un modelo de toma de decisiones (Wolny, 2014)

Recorrido del Cliente	Modelo de toma de decisiones
Involucra todos los puntos importantes y canales que comprometen al cliente con su proceso de compra.	Tiene una estructura totalmente jerarquica para que el cliente tome la decisión de comprar.
No es una estructura lineal, por lo general involucra distintos tipos de decisiones que desencadenan distintas consecuencias.	Estructura lineal
Ocurre debido a reflejos cognitivos, emocionales y de comportamiento.	Solo reflejos lógicos.
Sucedee normalmente en el marketing digital.	Sucedee en tiendas físicas.

2.2. Regresión Logística Binaria

En el trabajo desarrollado por Sperandei se afirma que, una regresión logística modela la probabilidad de un resultado basado en características individuales (Sperandei, 2013). Debido a que la probabilidad es una razón, lo que realmente se modela es el logaritmo de la probabilidad:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_mx_m \quad (1)$$

Donde π indica la probabilidad de un evento y los β_i son los coeficientes de regresión asociados con las variables explicativas x_i . Los β_i representan el cambio que se dará en el log $(\pi/(1-\pi))$ para una unidad de cambio en las variables independientes siempre que se mantengan constantes las demás variables independientes presentes en el modelo. Para una interpretación más simple el modelo también se escribe de la siguiente manera:

$$\text{Prob}(\pi) = \frac{e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_mx_m}}{1 + e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_mx_m}} \quad (2)$$

- La ecuación $e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_mx_m}$ representa a las variables independientes expresadas en la escala logit, en lugar del formato lineal original.
- La razón de esta transformación es para evitar valores de π fuera del intervalo $[0, 1]$.

Según Stoltzfus, la regresión logística es una forma eficaz y poderosa de analizar el efecto de un grupo de variables independientes en un resultado binario cuantificando la contribución

única de cada variable independiente. Mediante el uso de componentes de regresión lineal reflejados en la escala logit, la regresión logística identifica iterativamente la combinación lineal más fuerte de variables con la mayor probabilidad de detectar el resultado observado. Al validar modelos de regresión logística, hay numerosos métodos entre los que elegir, cada uno de que puede ser más o menos apropiado dependiendo de parámetros de estudio como el tamaño de la muestra (Stoltzfus, 2011). Para establecer validez interna, o confirmación de los resultados del modelo dentro del mismo conjunto de datos, los métodos comunes son:

- 1) El método de retención o dividir la muestra en dos y así calificar subgrupos antes de la construcción del modelo, con el grupo de "entrenamiento" utilizado para crear el modelo de regresión logística y el grupo de "prueba" utilizado para validarlo.
- 2) El método de validación cruzada o división de la muestra en un k número de subgrupos (o pliegues) separados y de igual tamaño para fines de construcción y validación de modelos.
- 3) El método "dejar uno fuera ", que es una variante del enfoque de el método anterior en el que el número de pliegues es igual a el número de sujetos en la muestra.
- 4) El método de diferentes formas de bootstrapping (es decir, obtener sub-muestras con reemplazo de todas las muestras).

Stoltzfus también indica que para la interpretación de los resultados de las variables individuales, éstas generalmente se presentan como razones de probabilidad o también denominados como odd ratios (OR). Los OR revelan la fuerza de la contribución de la variable independiente al resultado y se definen como $\frac{\pi}{1-\pi}$ para cada variable independiente. La relación entre el OR y la estimación del parámetro independiente β_i se expresa como $OR = e^{\beta_i}$. Establecido en esta fórmula, un cambio de 1 unidad en la variable independiente multiplica las probabilidades del resultado por la cantidad contenida en e^{β_i} . La interpretación de los OR también depende de si la variable independiente es continua o

categoría. En el caso de las variables continuas, primero se debe identificar una unidad de medida significativa para expresar mejor el grado de cambio en el resultado asociado con ese resultado independiente (Stoltzfus, 2011).

El investigador Ayer hace una comparación entre la regresión logística y el análisis de redes neuronales (Ayer, 2010). Esta comparación se presenta en la tabla 1.

Tabla 2. Comparación entre la regresión logística y las redes neuronales (Ayer, 2010)

Parámetro	Regresión Logística	Análisis de redes neuronales
Construcción del modelo	Requiere conocimiento estadístico	Requiere poco conocimiento estadístico
Habilidad para detectar relaciones complejas	Se le dificulta un poco	Modela automáticamente y es más fácil encontrar relaciones complejas
Habilidad para detectar interacciones	Requiere un modelo explícito de las interacciones	Puede detectar interacciones implícitas
Sobreajuste	Es poco propenso al sobreajuste	Propenso a tener problemas de sobreajuste
Habilidad discriminativa	Buena en general	Buena en general
Tiempo de proceso	Poco tiempo requerido	Tiempo regular
Compartir el modelo con otras personas	Fácil de compartir	Difícil de compartir
Intervalos de confianza	Fácil de calcular	Difícil de calcular
Facilidad para poder interpretar y tomar decisiones	Fácil de identificar predictores importantes	Es una caja negra difícil de identificar

Aunque el análisis de redes neuronales es una potente herramienta para predecir datos es demasiado complejo integrar esta solución en las herramientas de negocios usadas por la mayoría de empresas. Por otro lado, el modelo de regresión logística es una herramienta que obtiene prácticamente los mismos resultados con una menor complejidad a la hora de interpretar los resultados. Se debe comentar además que la técnica de regresión logística binaria es utilizada por el área de ventas digitales para realizar pronósticos de morosidad en

sus clientes, lo que genera una confianza mucho mayor al momento de proponer una solución analítica con el mismo equipo mencionado anteriormente (ventas digitales).

2.3. Técnica de Sobremuestreo

La investigación desarrollada por Mishra señala que cuando se analizan conjuntos de datos desequilibrados, los algoritmos de aprendizaje automático enfrentan dificultades. Las predicciones realizadas están sesgadas y tienen una precisión engañosa. Esto se debe a la falta de información sobre la clase minoritaria. Los algoritmos de aprendizaje automático suponen que los conjuntos de datos están equilibrados con pesos de clase iguales y, por lo tanto, tienden a clasificar cada muestra de caso de prueba en la clase mayoritaria para mejorar la métrica de precisión (Mishra, 2017).

El tema de sobremuestreo fue investigado por Fernández, quien plantea que el algoritmo SMOTE lleva a cabo un enfoque de sobremuestreo para reequilibrar el conjunto de entrenamiento original. En lugar de aplicar una réplica simple de las instancias de clases minoritarias, la idea clave de SMOTE es introducir ejemplos sintéticos. Estos nuevos datos se crean por interpolación entre varias instancias de clases minoritarias que se encuentran dentro de un vecindario definido (Fernandez, 2018). En la figura 2, se observa el conjunto de datos original con los grupos desequilibrados y el mismo conjunto de datos después de aplicar el algoritmo SMOTE. Nótese el incremento del número de observaciones en el grupo minoritario.

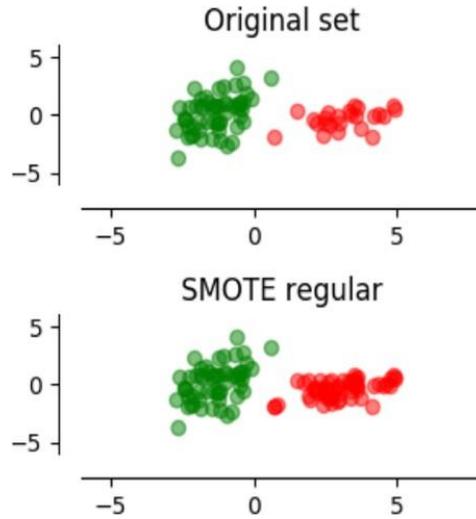


Figura 2. Conjuntos de datos original y de datos después de aplicar SMOTE (Fernandez, 2018)

El algoritmo SMOTE calcula la distancia entre los puntos de entrenamiento de la clase minoritaria para definir un vecindario, a partir del cual se seleccionan ejemplos para la creación de nuevos puntos sintéticos. Estas distancias generalmente se calculan usando la distancia euclidiana. De acuerdo a Maldonado el algoritmo SMOTE utiliza la distancia de Minkowski en lugar de la distancia euclidiana tradicional, explorando $q \in \{1, 2, \infty\}$, si $q = 1$ es equivalente a la distancia de Manhattan, si es 2 a la Euclidiana, pero también es aplicable para un numero mayor dependiendo del tipo de coordenadas en el plano cartesiano donde se evalúa el algoritmo. La distancia propuesta entre dos muestras de la clase minoritaria i e i' es la siguiente:

$$\left(\sum_{j \in S^t} |x_{i,j} - x_{i',j}|^q \right)^{\frac{1}{q}} \quad (3)$$

Dónde $x_{i,j}$ es la observación de la clase minoritaria y S es la cantidad total de la clase minoritaria de la muestra. Se debe tener en cuenta que $q \geq 1$ hasta S (Maldonado, 2017).

2.4. AUC

Según Hoo, el área bajo la curva ROC (AUC) es una medida global de la capacidad de una prueba para discriminar si una condición específica está presente o no. Un AUC de 0,5 representa una prueba sin capacidad de discriminación (es decir, no mejor que el azar) mientras que un AUC de 1,0 representa una prueba con discriminación perfecta. Al seleccionar un umbral óptimo (o punto de corte), se debe considerar los objetivos de la prueba de diagnóstico, considerando la importancia y los costos de una interpretación de falso positivo o falso negativo. (Hoo, 2017).

El área bajo la curva ROC, el AUC, se puede expresar de la siguiente forma:

$$AUC = \int_0^1 F_0(F_1^{-1}(1 - S_e)) dS_e \quad (4)$$

Dónde F_1 y F_0 representan funciones de distribución acumulativas y S_e es la sensibilidad, o tasa de verdaderos positivos que es la probabilidad de que un caso verdadero se clasifique correctamente, otro concepto es la especificidad nos dice qué porcentaje de casos falsos fue clasificado correctamente (Hand, 2010).

Según Jiménez la curva de característica operativa (ROC) puede tomar distintas formas siendo perpendicular con un área de 0.5 lo que significaría sin discriminación, o como se muestra en la figura 3 una curva con valor mayor a 0.5 que muestra que si existe una discriminación entre los valores analizados (Jiménez, 2012).

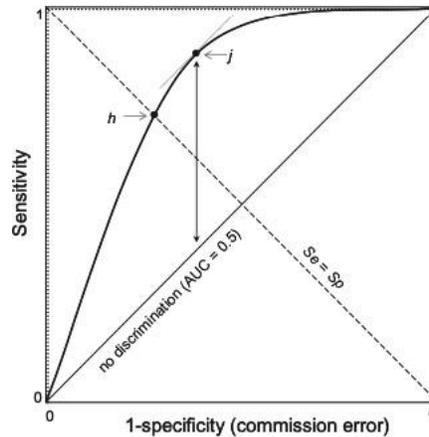


Figura 3 Curva AUC. Fuente: (Jiménez, 2012)

Para tomar una decisión debes entender estos conceptos: La sensibilidad se calcula dividiendo los casos positivos clasificados correctamente sobre los casos positivos clasificados correctamente y los casos verdaderos clasificados erróneamente, la especificidad se calcula dividiendo los casos falsos clasificados correctamente sobre los casos falsos clasificados correctamente y los casos falsos clasificados erróneamente. Si identificar correctamente los positivos es importante para nosotros, entonces deberíamos elegir un modelo con mayor Sensibilidad. Sin embargo, si identificar correctamente los negativos es más importante, entonces deberíamos elegir la especificidad como métrica de medición (Hand, 2010).

Por lo tanto, escoger el AUC como métrica generalizada para medir la eficiencia de un modelo estadístico es mayormente utilizado porque toma en consideración la sensibilidad dentro de un grupo evaluado. Tomar el valor correcto de AUC queda a juicio del investigador, siendo mencionado una buena practica tomar aquellos con un valor mayor a 0.7 (Hand, 2010).

2.5. Herramientas para la recolección y almacenamiento de datos

2.5.1 Web scraping

El "Web scraping" (también llamado "recolección web", "extracción de datos web" o incluso "minería de datos web"), se puede definir como "la construcción de un agente para descargar, analizar y organizar datos de la web en un de manera automatizada ". O, en otras palabras: en lugar de que un usuario humano manualmente copie el texto de una Web y lo pegue en una hoja de cálculo, el web scraping descarga esta tarea en un programa informático que puede ejecutarla mucho más rápido y más correctamente que un humano (Broucke, 2018).

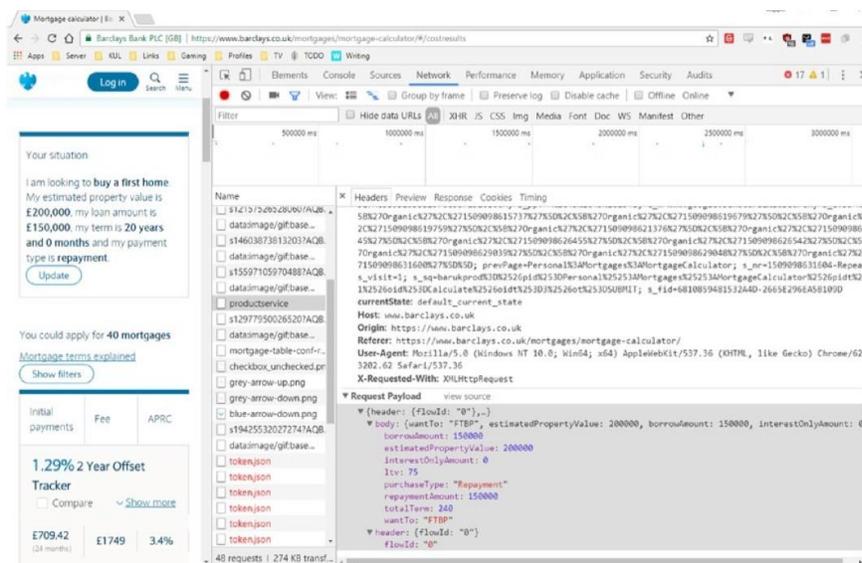


Figura 4 Ejemplo de Web scrapping de una web (Broucke, 2018)

2.5.2 Google Big Query

La computación en nube es un modelo que implementa un acceso conveniente y omnipresente a los recursos a los que se puede acceder con una interacción mínima con el proveedor de servicios. Los recursos pueden ser diversos, como infraestructura, servidores, almacenamiento, aplicaciones o servicios (Zinchenko, 2017). Las principales características del cloud computing son las siguientes:

- a) En demanda auto servicio. El tiempo del servidor y las capacidades de almacenamiento se aprovisionan automáticamente cuando es necesario sin interacción humana.

- b) Amplio acceso a la red. Los servicios y la infraestructura están disponibles a través de protocolos estándar en las redes.
- c) Puesta en común de recursos.
- d) Rápida elasticidad. En el caso de un rápido aumento en la demanda del servicio necesario escalará tanto como lo solicite el usuario.
- e) Servicio medido. Cada recurso puede ser controlado y monitoreado y el usuario se facturará según el uso del servicio.

BigQuery es una herramienta sin servidor para acceder rápidamente a los datos necesarios en el almacén de datos. Una base de datos típica está limitada por la velocidad del disco incluso si una consulta se ejecuta en paralelo. BigQuery almacena los datos en la nube en los diferentes centros de datos del mundo. También replica los clústeres existentes en diferentes ubicaciones, por lo que en el caso de una emergencia en uno de los centros de datos, los usuarios siempre tendrán acceso a los datos. BigQuery no es un lenguaje de consulta estructurado (SQL) ni una base de datos NoSQL por esa razón no puede ser usada para reemplazar un datawarehouse donde se puede alterar en cualquier momento mediante una consulta la data línea a línea (Zinchenko, 2017).

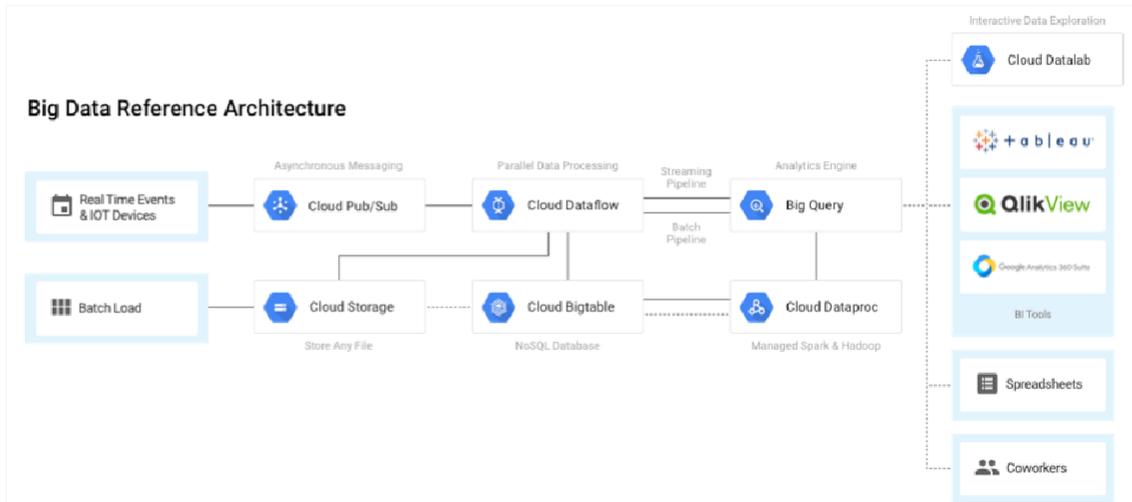


Figura 5 Arquitectura del almacenamiento de datos en la nube donde está Big Query (Zinchenko, 2017)

III. METODOLOGÍA

En este capítulo se presentan las diferentes etapas de la metodología del trabajo de monografía para alcanzar el objetivo general y los específicos.

3.1. Alcance

Este trabajo de investigación se realizó en un periodo de 30 días y los resultados de este trabajo solo se podrán aplicar para los datos utilizados. La razón de esta afirmación es que generalmente la tendencia de las visitas a la página Web es estacional y porque la mayoría de campañas que se realizan se repiten cíclicamente por los “cyber” dónde se lanzan muchas ofertas que atraen una mayor cantidad de visitas en unos días.

3.2. Tipo de investigación

El tipo de investigación del presente trabajo de monografía es no experimental y transaccional, ya que los datos se recolectaron en un solo momento. Además, la investigación es correlacional porque se quiere medir el grado de relación entre las variables independientes con la dependiente. Ya que uno de los objetivos es conocer si estas variables independientes contribuyen a que la probabilidad de que el usuario haga una transacción aumente o disminuya.

3.3. Población

La Web que se trata en todo el trabajo de monografía contiene varias páginas de destino mediante las cuales los usuarios pueden realizar una visita; sin embargo, solo se tomará en cuenta aquellas dónde los usuarios puedan realizar transacciones ya que el objetivo principal es describir los segmentos basados en que si realizan o no una transacción. Por este motivo, la población objetivo estará conformada por todos los usuarios que visitan las páginas “comerciales” de la Web del cliente durante un periodo de 30 días debido a la estacionalidad y por propia decisión del cliente.

3.4.Muestra

La muestra aleatoria convencional no se utilizó porque se utilizan todos los datos de la población mencionada en 3.3.

3.5.Recolección y modelamiento de datos

Se recolectaron los datos automáticamente cada vez que un usuario visito la Web mediante una herramienta de web scraping, estos datos serán almacenada en Google Bigquery y exportada a Python. El proceso de exportación de la data desde Google Bigquery a Python se logra mediante una API que provee Google llamada "Bigquery Storage". Se debe tener en cuenta que para que funcione correctamente esta API debe tenerse cargado e instalado la librería "pandas", debido a que la data importada es un marco de datos para lo cual "pandas" ayuda en manipularla.

El modelamiento se realizó usando la librería "sklearn" de Python (Python, 2020) y la data de entrenamiento será el 70% del total y la data de prueba el 30%, se realizo en una sola repetición. La librería "sklearn" proporciona muchas técnicas estadísticas de maquinas de aprendizaje para el presente trabajo de monografía se usará la correspondiente a regresión logística.

3.6. Variables

Para este trabajo de suficiencia profesional se utilizó la regresión logística donde las variables son las siguientes:

3.6.1. Variable Dependiente: Los valores de esta variable Y fueron:

- a) El usuario hace una transacción
- b) El usuario no hace una transacción

3.6.2. Variables Independientes:

- X1: Tiempo de navegación
- X2: Canales digitales:
 - a) Optimización de motores de búsqueda (Seo)
 - b) Posicionamiento en Google por medios pagados (Campanas)
 - c) Redes sociales
 - d) Email
 - e) Referencia
 - f) Directo
 - g) Otros canales.
 - h) (Other)
- X3: Tipo de dispositivo:
 - a) Desktop
 - b) Móvil
 - c) Tablet
- X4: Tipo de sistema operativo:
 - a) Windows phone
 - b) Android
 - c) iOS
 - d) Windows
- X5: Página de destino de la Web de Movistar:

- a) Home principal
 - b) Hogar Dúos
 - c) Hogar Tríos
 - d) Hogar internet
- X6: Usuario Nuevo
 - X7: Número de visitas

3.7.Análisis

Se realizó un análisis exploratorio utilizando gráficos de barras de las variables para preliminarmente establecer que las variables escogidas como independientes eran las correctas.

En cuanto a la técnica de Sobremuestreo SMOTE, como se mencionó anteriormente, esta es una metodología usada para que la predicción no sea sesgada y no provoque un error en la predictibilidad, está técnica crea pequeñas muestras sintéticas y se utiliza solo sobre los datos de entrenamiento para corregir el posible desbalanceo que pudiera existir.

Una vez realizado lo anterior, se procedió a generar un modelo de regresión logística que sirva principalmente para cumplir los objetivos específicos. El motivo de haber elegido la regresión logística es por la facilidad en la interpretación de sus coeficientes y por su sencillez en su aplicación. Además, porque ayuda a determinar las variables que son significativas, permitiendo identificar el aumento o disminución de la probabilidad de realizar una transacción en la Web, también ayuda a calcular la probabilidad de cada visita que realiza un usuario a la Web del cliente desde el momento en que se implemente este proyecto de la empresa consultora. Finalmente, para medir la predictibilidad del modelo se utilizó una matriz de confusión y el coeficiente del área bajo la curva (AUC) en la data de prueba. Si se obtenía un área con un umbral lo suficientemente aceptable tener en cuenta que el AUC aceptable será

una recomendación hecha por nosotros como expertos para posteriormente ser aprobada por el cliente, se procederá a trabajar con el modelo realizado anteriormente.

Con la evidencia estadística obtenida con la aplicación de la metodología anteriormente descrita se estableció un nivel de certeza de que el cálculo de la probabilidad era lo suficientemente confiable para proceder a realizar la segmentación. Por decisión de la empresa solo se establecieron tres segmentos ya que para cada segmento identificado se realizó un trabajo de personalización que requiere de una inversión monetaria por parte del cliente. Para hacer que los segmentos sean suficientemente simétricos se evaluó la posición de estos datos ordenados de menor a mayor probabilidad. Se usarán los percentiles 50 y 75 como puntos de corte para segmentar.

IV.

RESULTADOS Y DISCUSIÓN

El desarrollo de este modelo de regresión logística contribuyó a solucionar el problema de la deficiencia en la inversión en canales digitales que tiene el cliente y brindará experiencia a la empresa consultora en este tipo de proyectos, ya que esta tratando de abrir un nuevo servicio de aprendizaje automático. Esto fue valorado por el cliente quien tiene mucha confianza con la empresa consultora. Los resultados que se mostrarán a continuación se realizaron en el lenguaje de programación Python. La codificación fue realizada con el editor de texto “Sublime Text” debido a que contiene diversas opciones para mejorar el manejo y edición de gran cantidad de códigos. Para el procesamiento se usó una laptop con el sistema operativo Mac Os convirtiéndose en la terminal para el ingreso de los comandos de Python.

4.1. Resultados

4.1.1. Cargar las Librerías

Para realizar el modelo de regresión logística se tuvo que cargar las siguientes librerías de Python:

- Pandas: para cargar la data y realizar y cambio de nombres a las columnas.
- Numpy: para calcular estadísticas descriptivas como los percentiles.
- Sklearn: para aplicar algoritmos de aprendizaje automático. En este trabajo se usará para realizar el modelo de regresión logística, la tecnica de sobremuestreo SMOTE, la matriz de confusión y hallar el area bajo la curva (AUC).
- Matplotlib: para realizar gráficos durante el presente análisis que serán graficos de barras se debe tener en cuenta que esta librería funciona junto a Numpy.

```
import pandas as pd
import numpy as np
from sklearn import preprocessing
import matplotlib.pyplot as plt
plt.rc("font", size=14)
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
import seaborn as sns
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
```

Figura 6. Código de Python para cargar las librerías Pandas, Numpy, Sklearn y Matplotlib

En la figura 7, se observa que la proporción de usuarios que visitan la web y hacen transacciones es de aproximadamente un 20% del total por otro lado los que no lo hacen es del 80% aproximadamente. Esto muestra un total desbalance de las clases siendo la clase minoritaria los usuarios que realizan transacciones.

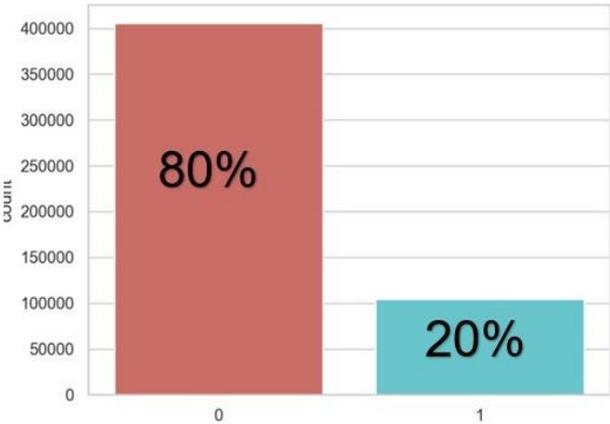


Figura 7. Cantidades de usuarios que realizan transacciones (1) y que no realizan transacciones (0)

4.1.2. Análisis univariante

Se observa en la Figura 8 que, para la variable “canal” se tiene más cantidad de pedidos en “Campanas” y “Seo”. En los demás canales la cantidad de transacciones que se han realizado es mucho menor.

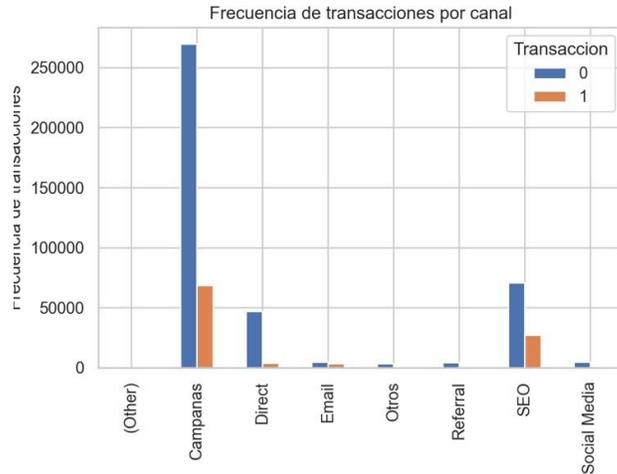


Figura 8. Análisis univariante de la variable “canal”

En la figura 9, para la variable “tipo de dispositivo” se observa que las transacciones no son iguales en todas las categorías, se tiene menos pedidos en Mobile comparado con el resto.

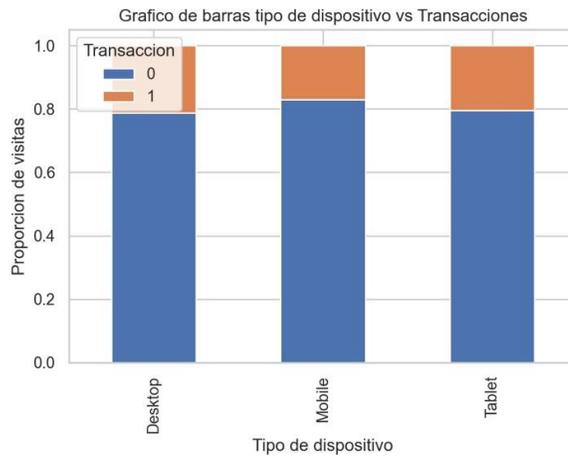


Figura 9. Análisis univariante de la variable “tipo de dispositivo”

En la figura 10, para las variables referentes al tipo de sistema operativo que tienen los usuarios se observa que existe diferencias evidentes entre la proporción de usuario que hicieron una transacción frente a los que no hicieron una transacción entre las categorías, siendo Windows donde los usuarios hacen más transacciones comparado con los demás sistemas operativos.

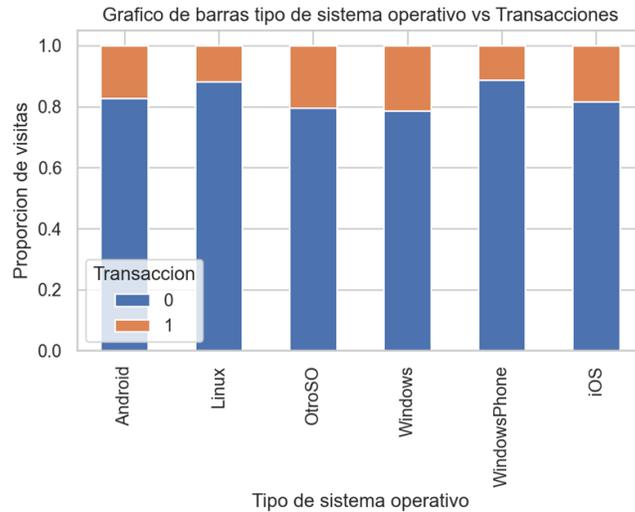


Figura 10. Análisis univariante de la variable “tipo de sistema operativo”

En la figura 11, se observa que en la variable “Usuario Nuevo” los usuarios que no entran por primera vez a la web su proporción de transacciones frente a los que no lo hacen es mucho mayor a que si fuera la primera vez que entran a la Web.

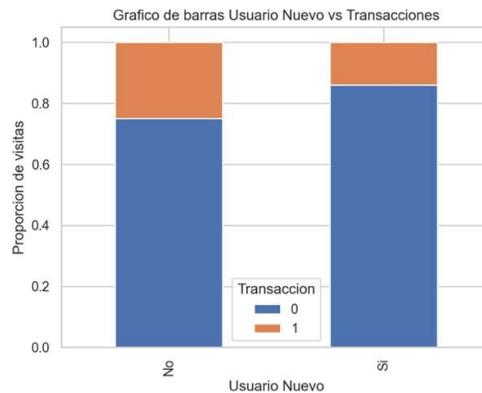


Figura 11. Análisis univariante de la variable “Usuario Nuevo”

En la figura 12, se observa que en la variable “Pagina de destino” es en “home_hogar” donde los usuarios tienen una mayor proporción de transacción frente a los que no.

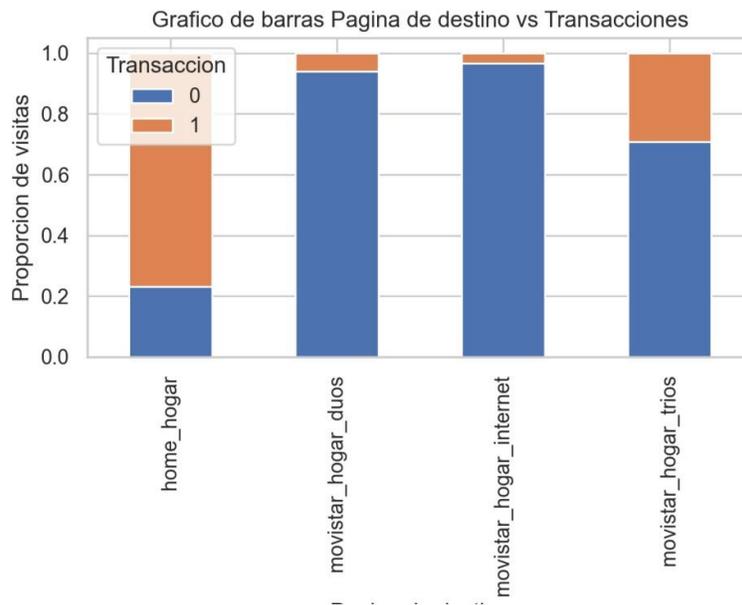


Figura 12. Análisis univariante de la variable “Página de destino”

4.1.3. Categorización de las variable cualitativas

Existen variables como “canal”, “tipo de dispositivos”, “tipo de sistema operativo” y “página de destino” que necesitan ser codificadas para volverse variables ficticias donde cada categoría de estas variables será una variable independiente binaria lo que permitirá una mejor interpretación y cálculos más sencillos para las razones de probabilidad.

```

cat_vars=['canal','tipo_dispositivo','tipo_so','landing_page']
for var in cat_vars:
    cat_list='var'+ '_' +var
    cat_list = pd.get_dummies(data[var], prefix=var)
    data1=data.join(cat_list)
    data=data1

cat_vars=['canal','tipo_dispositivo','tipo_so','landing_page']
data_vars=data.columns.values.tolist()
to_keep=[i for i in data_vars if i not in cat_vars]

data_final=data[to_keep]
data_final.columns.values
data_final.head()

```

Figura 13. Código en lenguaje de programación Python para categorizar las variables cualitativas

```

>>> data_final=data[['Transaccion','menos_30','de_30_60','de_60_mas','nuevaVisita','canal_Campanas', 'canal_Direct',
...
'canal_Email', 'canal_Otros', 'canal_Referral', 'canal_SEO',
...
'canal_Social Media', 'tipo_dispositivo_Desktop',
...
'tipo_dispositivo_Mobile', 'tipo_dispositivo_Tablet',
...
'tipo_so_Android', 'tipo_so_Linux', 'tipo_so_OtrosSO',
...
'tipo_so_Windows', 'tipo_so_WindowsPhone', 'tipo_so_IOS',
...
'landing_page_home_hogar', 'landing_page_movistar_hogar_duos',
...
'landing_page_movistar_hogar_internet',
...
'landing_page_movistar_hogar_trios']]
>>> data_final.head()
Transaccion menos_30 de_30_60 de_60_mas ... landing_page_home_hogar landing_page_movistar_hogar_duos landing_page_movistar_hogar_internet landing_page_movistar_hogar_trios
0 1 1 0 0 ... 0 0 0
1 1 1 0 1 0 ... 0 0 0
2 1 0 0 1 ... 0 1 0
3 1 0 0 1 ... 0 1 0
4 1 0 0 1 ... 0 0 0
1

```

Figura 14. Data con las nuevas variables binarias creadas

En el caso de la variable “número de usuarios” se opto por no realizar una discretización de estos a pesar de ser una variable numérica, se limito el seguimiento de usuario a un máximo de 5 visitas al día, los motivos fueron dos: el primero es que se paga por GB de espacio y cada línea es una visita web en nuestra base de datos y la segunda razón es que google no permite guardar las cookies de un mismo usuario por mas de 24 horas. Al ser un rango limitado de visitas no hay necesidad de discretizar.

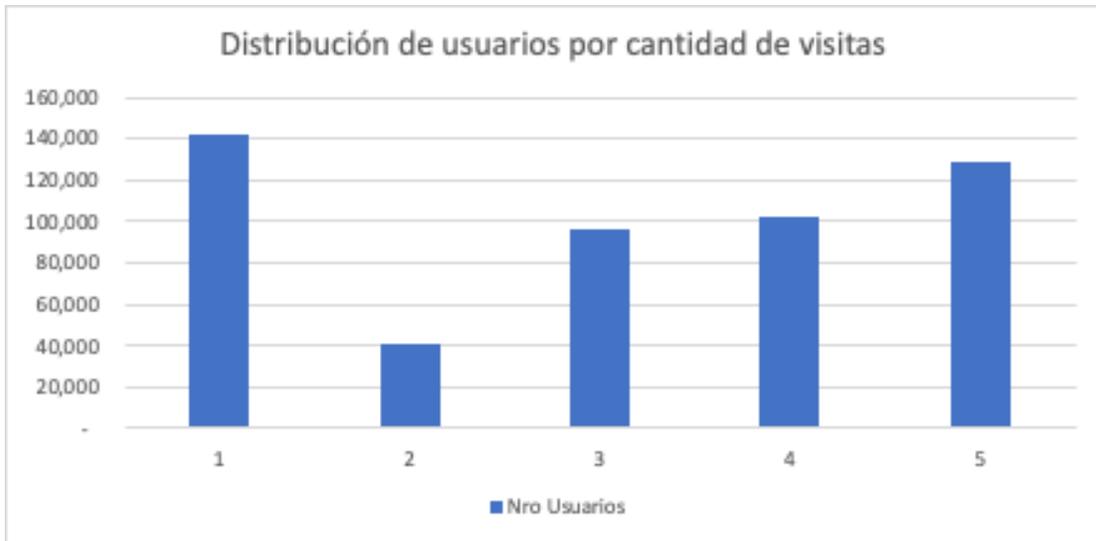


Figure 15 Distribución de usuarios por nro de visitas

4.1.4. Técnica de sobremuestreo

Se aplica la técnica de sobremuestreo “SMOTE” sobre los datos de entrenamiento, para solucionar el desbalanceo entre las clases, además se dividen los datos de entrenamiento y de prueba.

```
X = data_final.loc[:, data_final.columns != 'Transaccion']
y = data_final.loc[:, data_final.columns == 'Transaccion']

from imblearn.over_sampling import SMOTE
os = SMOTE(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
columns = X_train.columns
os_data_X,os_data_y=os.fit_sample(X_train, y_train)
os_data_X = pd.DataFrame(data=os_data_X,columns=columns )
os_data_y= pd.DataFrame(data=os_data_y,columns=['Transaccion'])
```

Figura 16. Código en lenguaje de programación Python dónde se aplicará la técnica de sobremuestreo “Smote”

Se observa en la figura 17 que ahora los usuarios que hicieron una transacción son la misma cantidad que los que no hicieron una transacción.

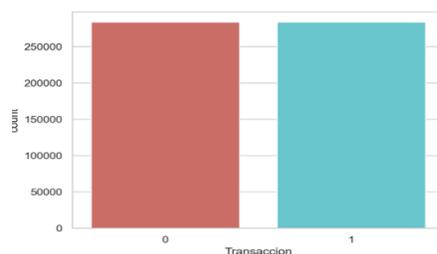


Figura 17. Clases balanceadas después de aplicar la técnica de sobremuestreo

4.1.5. Modelo de regresión logística

La figura 18 presenta el código para realizar la regresión logística. Mientras que la figura 15 presenta el modelo de regresión logística obtenido con Python. Nótese que los coeficientes de las variables “Android”, “Windows” y “iOS” no contribuyen a predecir si el usuario hará o no una transacción en la Web.

```

from sklearn.linear_model import LogisticRegression
from sklearn import metrics
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

```

Figura 18. Código en lenguaje de programación Python para realizar un modelo de regresión logística a la data de entrenamiento

Results: Logit					
Model:	Logit	Pseudo R-squared:	0.204		
Dependent Variable:	Transaccion	AIC:	626230.1779		
Date:	2020-10-30 16:10	BIC:	626488.9154		
No. Observations:	567760	Log-Likelihood:	-3.1309e+05		
Df Model:	22	LL-Null:	-3.9354e+05		
Df Residuals:	567737	LLR p-value:	0.0000		
Converged:	1.0000	Scale:	1.0000		
No. Iterations:	6.0000				
	Coef.	Std.Err.	z	P> z	[0.025 0.975]
nuevaVisita	-0.8643	0.0067	-128.2647	0.0000	-0.8776 -0.8511
menos_30	3.9233	0.4581	8.7167	0.0000	3.0411 4.8054
de_30_60	4.4575	0.4503	9.8984	0.0000	3.5749 5.3401
de_60_mas	4.0904	0.4500	9.0890	0.0000	3.2084 4.9725
SEO	-1.8884	0.1784	-10.5878	0.0000	-2.2380 -1.5389
Campana	-0.6434	0.1783	-3.6077	0.0003	-0.9929 -0.2939
SocialMedia	-2.3786	0.1822	-13.0581	0.0000	-2.7356 -2.0216
Email	-0.5319	0.1796	-2.9612	0.0031	-0.8840 -0.1799
Referencia	-2.7931	0.1823	-15.3242	0.0000	-3.1503 -2.4358
Directo	-2.3847	0.1787	-13.3430	0.0000	-2.7350 -2.0344
OtroCanal	-3.0963	0.1836	-16.8647	0.0000	-3.4561 -2.7364
Mobile	-1.1630	0.3362	-3.4597	0.0005	-1.8218 -0.5041
Desktop	-0.9272	0.3322	-2.7915	0.0052	-1.5783 -0.2762
Tablet	-0.9110	0.3363	-2.7088	0.0068	-1.5702 -0.2518
Android	-0.0196	0.0709	-0.2768	0.7819	-0.1586 0.1194
Windows	0.0180	0.0219	0.8202	0.4121	-0.0250 0.0609
iOS	-0.0164	0.0728	-0.2248	0.8222	-0.1590 0.1263
Linux	-0.4025	0.0511	-7.8698	0.0000	-0.5027 -0.3022
WindowsPhone	-0.5853	0.1541	-3.7981	0.0001	-0.8073 -0.2833
home_hogar	1.6091	0.2672	6.0233	0.0000	1.0855 2.1327
movistar_hogar_duos	-2.9285	0.2628	-11.1435	0.0000	-3.4435 -2.4134
movistar_hogar_trios	-0.9491	0.2627	-3.6130	0.0003	-1.4640 -0.4342
movistar_hogar_internet	-3.6841	0.2633	-13.9925	0.0000	-4.2001 -3.1680

Figura 19. Resultados del modelo de regresión logística con todas las variables de la data de entrenamiento

En el nuevo modelo, luego de eliminar las variables no significativas, se observa que todas las variables son significativas y todas aportan información relevante para predecir la probabilidad de que el usuario hagan una transacción cuando entra a la web.

Results: Logit						
Model:		Logit	Pseudo R-squared:		0.204	
Dependent Variable:		Transaccion	AIC:		626225.0933	
Date:		2020-10-30 16:15	BIC:		626450.0824	
No. Observations:		567760	Log-Likelihood:		-3.1309e+05	
Df Model:		19	LL-Null:		-3.9354e+05	
Df Residuals:		567740	LLR p-value:		0.0000	
Converged:		1.0000	Scale:		1.0000	
No. Iterations:		6.0000				
	Coef.	Std.Err.	z	P> z	[0.025	0.975]
nuevaVisita	-0.8643	0.0067	-128.2826	0.0000	-0.8776	-0.8511
menos_30	3.9320	0.4492	8.7531	0.0000	3.0516	4.8125
de_30_60	4.4662	0.4495	9.9367	0.0000	3.5853	5.3471
de_60_mas	4.0991	0.4492	9.1258	0.0000	3.2187	4.9795
SEO	-1.8876	0.1784	-10.5819	0.0000	-2.2373	-1.5380
Campana	-0.6426	0.1784	-3.6029	0.0003	-0.9922	-0.2930
SocialMedia	-2.3781	0.1822	-13.0539	0.0000	-2.7352	-2.0210
Email	-0.5310	0.1797	-2.9560	0.0031	-0.8832	-0.1789
Referencia	-2.7924	0.1823	-15.3185	0.0000	-3.1496	-2.4351
Directo	-2.3840	0.1787	-13.3375	0.0000	-2.7343	-2.0337
OtroCanal	-3.0954	0.1836	-16.8582	0.0000	-3.4553	-2.7355
Mobile	-1.1914	0.3315	-3.5943	0.0003	-1.8410	-0.5417
Desktop	-0.9189	0.3314	-2.7727	0.0056	-1.5684	-0.2694
Tablet	-0.9352	0.3324	-2.8139	0.0049	-1.5866	-0.2838
Linux	-0.4189	0.0465	-9.0106	0.0000	-0.5100	-0.3278
WindowsPhone	-0.5660	0.1371	-4.1274	0.0000	-0.8348	-0.2972
home_hogar	1.6089	0.2672	6.0217	0.0000	1.0852	2.1325
movistar_hogar_duos	-2.9287	0.2628	-11.1430	0.0000	-3.4439	-2.4136
movistar_hogar_trios	-0.9494	0.2627	-3.6137	0.0003	-1.4644	-0.4345
movistar_hogar_internet	-3.6843	0.2633	-13.9918	0.0000	-4.2004	-3.1682

Figura 20. Resultados del modelo de regresión logística solo con las variables que dan información significativa al modelo original

En la siguiente figura se muestra como se obtuvo las probabilidades para cada observación en la data de prueba. Se debe tener en cuenta que para la data de prueba no se utilizo el algoritmo de sobremuestreo SMOTE ya que se quiere evaluar la predictibilidad de la data en el escenario más real posible y es cuando las dos clases de la variable dependiente están desbalanceadas. Seguido por esto se obtiene el área bajo la curva lo que sirve para evaluar la predictibilidad del modelo.

```

y_pred = logreg.predict(X_test)
y_prob = logreg.predict_proba(X_test)[:,1]
print('Precisión del clasificador de regresión logística en el conjunto de prueba: {:.2f}'.format(logreg.score(X_test, y_test)))

```

Figure 21 Código en lenguaje de programación Python para obtener las probabilidades estimadas y obtener el área bajo la curva

En la siguiente tabla se menciona cada variable con una descripción de esta y si aporta positivamente o negativamente a la probabilidad de que el usuario que visita la Web haga una transacción o no.

Tabla 3. Descripción de todas las variables del modelo de regresión logística

Variable Independiente	Descripción	Signo del Coeficiente de Regresión
nuevaVisita	Es la primera visita del usuario a la Web	Negativo
menos_30	El tiempo de navegación en la Web fue de menos de 30 segundos	Positivo
de_30_60	El tiempo de navegación en la Web fue desde 30 segundos a menos de 60	Positivo
de_60_mas	El tiempo de navegación en la Web fue desde 60 segundos a más	Positivo
SEO	Optimización de motores de búsqueda	Negativo
Campana	Posicionamiento en google por medios pagados	Negativo
SocialMedia	Redes sociales	Negativo
Email	Email	Negativo
Referencia	Referenciado de una Web anterior a la actual	Negativo
Directo	El usuario entra la url por iniciativa propia	Negativo
OtroCanal	Un canal digital diferente a los mencionados	Negativo
Mobile	Tipo de dispositivo móvil	Negativo
Desktop	Computadora de escritorio	Negativo
Tablet	iPad	Negativo
Linux	Sistema operativo Linux	Negativo
WindowsPhone	Sistema operativo móvil de Windows	Negativo
home_hogar	La página donde se ofrecen todos los productos	Positivo
movistar_hogar_duos	La página donde se ofrecen los productos duos	Negativo
movistar_hogar_trios	La página donde se ofrecen los productos tríos	Negativo
movistar_hogar_inter net	La página donde se ofrece el producto de solo internet	Negativo

4.1.6. Validación del modelo

Se muestra la matriz de confusión:

```
>>> confusion_matrix = confusion_matrix(y_test, y_pred)
>>> print(confusion_matrix)
[[57853 26973]
 [13748 71754]]
```

Figura 22. Código en lenguaje de programación Python para obtener la matriz de confusión

Se muestra en la siguiente tabla 4 la cantidad de verdaderos positivos, verdaderos negativos y los errores tipo I y tipo II. Positivo es cuando se afirma que el usuario hara la transacción en la Web y negativo cuando no realizara la transacción. Los errores tipo I se dan cuando se clasifican como a usuarios que van a comprar en la Web pero en realidad no lo hicieron y el los errores Tipo II se dan cuando se predice que no realizaran una transacción en la Web pero realmente si lo hicieron.

Tabla 4. Matriz de confusión

MATRIZ DE CONFUSIÓN		Predictivo	
		Positivo	Negativo
Real	Positivo	57,853	26,973
	Negativo	13,748	71,754
Precisión		76%	

```

from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, logreg.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, logreg.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()

```

Figura 23. Código en lenguaje de programación Python para realizar la curva ROC

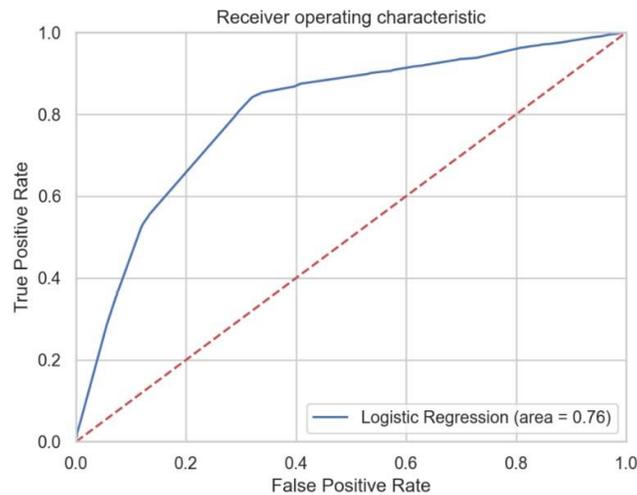


Figura 24. Curva ROC

Se tiene un AUC de 0.76 lo que significa que existe un 76% de probabilidades de que el modelo clasifique adecuadamente cada observación. La interpretación sería la siguiente: Existe un 76% de confianza que el usuario realice una transacción en la Web si el modelo de regresión logística binaria pronostica que se realizara la transacción.

4.1.7. Segmentación

Se observa que mediante un diagrama de cajas que la mayoría de probabilidades que resultaron de los datos de prueba están entre 30% y 70%. Esto ayuda a confirmar que

efectivamente la mayoría de los usuarios en los datos de prueba probablemente no hacen una transacción en la web lo cual es el comportamiento común de los mismos.

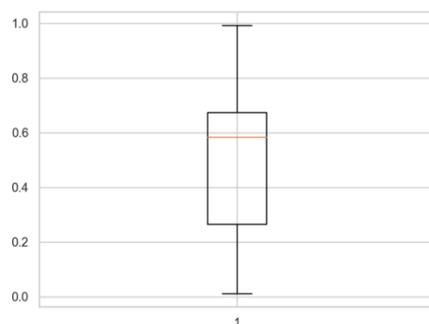


Figura 25. Diagrama de cajas y bigotes

Segmento de Alta probabilidad: Aquellos usuarios que visitaron la Web que estén posicionados después del percentil 75. Corresponden a los usuarios que tienen la probabilidad de hacer una transacción por encima del 67%.

Segmento de Media probabilidad: Aquellos usuarios que visitaron la Web que estén posicionados entre el percentil 50 y 75. Corresponden a los usuarios que tienen la probabilidad de hacer una transacción de 58% hasta 67%.

Segmento de Baja probabilidad: Aquellos usuarios que visitaron la Web que estén posicionados entre por debajo percentil 50: Corresponden a los usuarios que tienen la probabilidad de hacer una transacción menor a 58%.

```
>>> y_prob = logreg.predict_proba(X_test)[:,-1]
>>> np.percentile(y_prob, 50)
0.5838336764085321
>>> np.percentile(y_prob, 75)
0.6736791893731835
```

Figura 26. Código en lenguaje de programación Python para obtener los percentiles 50 y 75

En la tabla 5 se observa el criterio para la segmentación y presenta a todos los usuarios clasificados como “Baja Probabilidad” si tienen una probabilidad de realizar una transacción

menor a 0.58, como “Media Probabilidad” si su probabilidad es mayor o igual a 0.58 pero menor a 0.67 y finalmente como “Alta Probabilidad” si su probabilidad es mayor o igual a 0.67. Se espera una cantidad de usuarios mensuales que se distribuyan en estos segmentos de tal manera que sean los mismos para el segmento de alta y media probabilidad y del doble para el segmento de baja probabilidad.

Tabla 5. Segmentos e intervalo de probabilidades

Segmento	Intervalo de probabilidad de realizar una transacción	Total de clientes y proporción del total
Baja Probabilidad	[0,0.58>	254,865 y 50%
Media Probabilidad	[0.58,0.67>	127,433 y 25%
Alta Probabilidad	[0.67,1]	127,432 y 25%

En la tabla 6 se muestra un ejemplo de 3 observaciones de los datos de prueba donde se aplico la ecuación de la regresión logística binaria y se segmentaron.

Tabla 6 Ejemplo de tres observaciones a las que se les aplico el modelo estadístico

Variables	Coefficientes	O ₁	O ₂	O ₃
Nuevavisita	-128.2826	0	0	0
menos_30	8.75	1	0	0
de_30_60	9.936	0	1	1
de_60_mas	9.1258	0	0	0
SEO	-10.5819	0	1	0
Campana	-3.6029	1	0	0
SocialMedia	-13.0539	0	0	0
Email	-2.956	0	0	1
Referencia	-15.3185	0	0	0
Directo	-13.3375	0	0	0
OtroCanal	-16.8582	0	0	0
Mobile	-3.5943	1	0	0
Desktop	-2.7727	0	1	1
Tablet	-2.8139	0	0	0
Linux	-9.0106	0	0	0
Windowsphone	-4.1274	1	0	1

Home hogar	6.0217	1	1	0
movistar_hogar_duos	-11.143	0	0	0
movistar_hogar_trio	-3.6137	0	0	1
movistar_hogar_internet	-13.9918	0	0	0
$\frac{e^{\beta_0 + \beta_{ixi}}}{1 + e^{\beta_0 + \beta_{ixi}}}$		0.969144539	0.412679356	0.644190726

La interpretación es la siguiente:

- a) Cuando un usuario entra a la web y no es "usuario nuevo", su tiempo de navegación en la Web fue de menos de 30 segundos, entro mediante un canal de campaña de google, con navegador en modo mobile, con el sistema operativo Windowsphone y por la web de destino Home hogar, tiene 96% de probabilidades que realice una transacción y estará en el segmento de "alta probabilidad".
- b) Cuando un usuario entra a la web y no es "usuario nuevo", su tiempo navegación en la Web fue entre 30 y 60 segundos, entro mediante el canal orgánico SEO, con navegador en modo desktop, con el sistema operativo Android, por la web de destino Home Hogar, tiene 41% de probabilidades que realice una transacción y estará en el segmento de "baja probabilidad".
- c) Cuando un usuario entra a la web y no es "usuario nuevo", su tiempo navegación en la Web fue entre 30 y 60 segundos, entro mediante el canal email, con navegador en modo desktop, con el sistema operativo Windowsphone, por la web de destino Home hogar trio, tiene 64% de probabilidades que realice una transacción y estará en el segmento de "media probabilidad".

4.2.Discusión

Para obtener los resultados se trabajó con el lenguaje de programación Python, donde se cargaron las librerías correspondientes a la manipulación de la data y algoritmos de aprendizaje automático. Se optó por realizar un análisis univariante de cada variable independiente comparado con la variable dependiente para poder visualizar si efectivamente existe algún tipo de indicio de discriminación de estas. Efectivamente, se observó que existe tal poder discriminatorio como por ejemplo los usuarios que entran a la Web por la página de destino “Home Hogar” tienen una mayor proporción de pedido sobre visita frente a las demás.

Se adoptó la decisión de categorizar todas las variables cualitativas para una mejor interpretación del modelo de regresión logística. Esto permitió saber si el tipo de dispositivo, tipo de sistema operativo, página de destino o canal digital ayuda o no a predecir si el usuario realizara la transacción en la Web cada uno independientemente, se debe saber exactamente de cual. Como se observó inicialmente que existe un total desbalance entre las clases se aplicó la técnica de sobremuestreo SMOTE en los datos de entrenamiento para realizar el modelo de dos clases y estas estén balanceadas. Un conjunto de datos con las dos clases balanceadas en el grupo de entrenamiento y de prueba, ayuda a que el modelo pueda predecir mejor ya que se entrenara con la misma cantidad de usuarios en ambos grupos.

El modelo de regresión logística inicial indicó que las tres variables “Android”, “Windows” y “iOS”, no aportan información importante para predecir, se adoptó la decisión de retirarlas porque al hacerlo el modelo mejoró su predictibilidad y además redujo la cantidad de datos necesarios, algo que es muy importante ya que cliente paga por GB de la data que se extrae. Con el modelo final se obtienen las siguientes interpretaciones:

- Se observa que cuando el usuario es nuevo contribuye significativamente a la probabilidad que el usuario no haga una transacción con p-valor de 0.0000001.

- Los usuarios que navegan entre 30 o 60 segundos contribuyen significativamente a la probabilidad que el usuario haga una transacción con p-valor de 0.0000001.
- Los usuarios que visitan la Web mediante los canales Referencia, Directo y Social Media contribuyen significativamente a la probabilidad que el usuario haga una transacción con p-valores de 0.0000003, 0.0000001 y 0.000000001 respectivamente.
- Los usuarios que entran a la página de destino “Home_hogar” contribuyen significativamente a la probabilidad de que el usuario haga una transacción con p-valor de 0.00003.

Para confirmar el poder predictivo del modelo se aplicaron las técnicas de matriz de confusión y curva ROC para la data de prueba. La matriz de confusión mostró en una tabla la cantidad de usuarios que fueron clasificados correctamente, así como los que no lo fueron, por el modelo de regresión logística que construido. El AUC (área bajo la curva) fue de 76% lo cual significa que es “Óptimo”, no existe realmente una regla de que tan alto debería ser un AUC ya que depende mucho del objetivo. En cuanto al objetivo específico de categorizar a los usuarios en tres segmentos basados en la probabilidad que proporciona el modelo de regresión logística, se logró un 76% de acierto en que la clasificación sea correcta, lo que fue considerado como aceptable por el cliente.

Con las puntuaciones que se obtuvieron del modelo se realizó una segmentación, y se propuso hacerlo en tres segmentos bien diferenciados. Para este fin, se hizo un diagrama de cajas y bigotes para visualizar en que rango de puntuaciones se concentran la mayoría de los clientes. Se utilizaron los percentiles 50 y 75 ya que la cantidad de usuarios para los segmentos de alta y media probabilidad deben ser iguales y los usuarios de baja probabilidad sean el doble. Se destinará una mayor inversión para la baja probabilidad porque son esos

usuarios a quienes se debe persuadir a que terminen realizando una compra mediante una personalización en el sitio Web.

V. CONCLUSIONES

Como consecuencia de lo expuesto en este trabajo de monografía se obtuvo una segmentación apropiada para los usuarios que visitan la Web con sustento estadístico lo cual ayudó a la empresa consultora a poder abrir una nueva línea de negocios. Con los resultados obtenidos, se determinó cuáles fueron las variables que tienen un nivel de significancia al momento de aumentar o disminuir la probabilidad de que un usuario que entra a la Web logre realizar una transacción y así lograr categorizarlo en alguno de los tres segmentos. Se detalla a continuación en cada párrafo las conclusiones obtenidas de cada parte del proceso.

1. Desde el principio por decisiones del negocio se escogió que se harían solo tres segmentos porque se planeó realizar estrategias totalmente diferenciadas con cada uno de ellos. Se aplica en un marco temporal de 30 días por la estacionalidad del negocio y se usa la técnica estadística SMOTE para lograr eliminar la posible dispersión que generaría trabajar con datos no balanceados.
2. Los segmentos originados por las puntuaciones calculadas por el modelo de regresión logística binaria demostraron que se tiene base para aplicar al menos 3 estrategias diferentes. Siendo el segmento de baja probabilidad al cual se les diseñara Webs informativas, media probabilidad la Web principal y finalmente alta probabilidad donde se se les mostrara la tienda online directamente.
3. El modelo de regresión logística binaria identificó la significancia de cada variable independiente respecto a la variable dependiente. Por consiguiente, se puede interpretar que los usuarios categorizados como “Usuario nuevo” contribuyen a que el modelo tenga una baja probabilidad de hacer una transacción, cuando el usuario está entre 60 y 90 segundos en la Web aporta información positiva en la ecuación logística, se concluye que dentro del marco temporal de la visita de un usuario existe un rango donde aumenta la

probabilidad de que se decida en hacer la transacción, existen canales digitales como Referencia, Social Media o Directo que no tienen significancia con la variable dependiente y la página de destino “Home hogar” es donde los usuarios tendrían una mayor probabilidad de hacer una transacción.

VI. RECOMENDACIONES

Antes de finalizar, se sugieren algunas recomendaciones en base a los resultados y las conclusiones que se obtuvieron luego de realizar este trabajo de suficiencia profesional.

1. Se recomienda realizar nuevas técnicas estadísticas que se usarán con los segmentos descritos en el presente trabajo de monografía. Un ejemplo es usar técnicas de minería de texto en los comentarios que se recolecten para los usuarios que son clasificados como “baja probabilidad”, para encontrar las palabras claves de porque deciden no hacer una transacción, y así realizar estrategias de marketing digital más personalizadas.
2. Se aconseja tener en la data para futuros modelamientos la variable “canal_anterior” que se definiría como el canal digital de la visita anterior a la que se analiza en el momento de realizar el modelamiento. La inclusión de esta nueva variable podría hacer que el modelo de regresión logística sea más preciso; además que, con un análisis descriptivo de esta nueva variable se tendría una mejor visibilidad del comportamiento de los usuarios que visitan la Web y así poder tomar mejores decisiones.
3. Se sugiere implementar una nueva línea de negocio aplicando técnicas estadísticas para obtener soluciones analíticas a los problemas de análisis de datos de los clientes actuales y futuros. Este trabajo de suficiencia profesional permitió obtener un valioso aprendizaje y experiencia.
4. Se recomienda utilizar las técnicas de minería de texto para encontrar palabras claves en redes sociales, la técnica de análisis de supervivencia para redistribuir el presupuesto en canales digitales y las diferentes técnicas regresión lineal y no lineal para asignar los indicadores claves como el costo por click. Estos son algunos ejemplos del gran abanico de proyectos que se podrían aplicar a partir de ahora en la empresa consultora.

5. Se aconseja realizar diferentes modelos de regresión logística binarios para cada “Página de destino” de la Web. La finalidad es poder evaluar en un futuro cercano en cuales se necesita un mayor trabajo de mejora continua además de como, el modelo nos indico que solo una “Página de destino” (Home Principal) es una variable significativa para calcular la probabilidad de realizar una transacción.

VII. BIBLIOGRAFÍA

- Ayer, T. (2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation.
- Bayoude, K. (2018). How Machine Learning Potentials are transforming the Practice of Digital Marketing: State of the Art.
- Broucke, S. v. (2018). Practical Web Scraping for Data Science.
- Dolnicar, S. (2018). Market Segmentation Analysis.
- Fernandez, A. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary.
- Fonseca, J. (2011). Why does Segmentation Matter? Using Mixed Methodology to Identify Market Segments.
- Goyat. (2011). ALGO PROBANDO.
- Goyat, S. (2011). The basis of market segmentation: a critical review of literature.
- Grishikashvili, K. (2014). Investigation into Big Data Impact on Digital Marketing .
- Hand, D. J. (2010). Evaluating diagnostic tests: the area under the ROC curve and the balance of errors.
- Hoo, Z. H. (2017). What is an ROC curve?
- Jiménez, A. (2012). nsights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modellinggeb_683 498..507.
- Kovacs, G. (2019). smote-variants: a Python Implementation of 85 Minority Oversampling Techniques.
- Li, S. (2016). Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns.
- Maldonado, S. (2017). An alternative SMOTE oversampling strategy for high-dimensional datasets.

Mishra, S. (2017). Handling Imbalanced Data: SMOTE vs. Random Undersampling.

Muschelli, J. (2020). ROC and AUC with a Binary Predictor: a Potentially Misleading Metric.

Python. (2020). scikit-learn.

Saura, J. R. (2020). Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics.

Sperandei, S. (2013). Understanding logistic regression analysis. Rio de Janeiro.

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer.

Wolny, J. (2014). Mapping customer journeys in multichannel decision-making.

Zinchenko, Y. (2017). Big Data and Google Cloud Platform.