UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DISEÑO DE UN MODELO DE MACHINE LEARNING PARA MEJORAR LA VENTA DE RUTAS DIARIAS EN LA STARTUP DE LOGÍSTICA WARECLOUDS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL INDUSTRIAL

GABRIELA PAZ NEGRETE GODOY

PROFESOR GUÍA:
JUAN PABLO ROMERO GODOY

MIEMBROS DE LA COMISIÓN:
IVÁN DÍAZ CAMPOS
FELIPE VILDOSO CASTILLO

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: GABRIELA PAZ NEGRETE GODOY
FECHA: 2022
PROF. GUÍA: JUAN PABLO ROMERO GODOY

DISEÑO DE UN MODELO DE MACHINE LEARNING PARA MEJORAR LA VENTA DE RUTAS DIARIAS EN LA STARTUP DE LOGÍSTICA WARECLOUDS

Wareclouds es una startup que comenzó a operar en 2020 y ofrece servicios de logística a pequeñas y medianas empresas para sus productos vendidos en modalidad online. El servicio entregado se basa en un modelo colaborativo, donde los productos son almacenados y armados en casas de particulares llamados wareclouds y posteriormente los pedidos son entregados por repartidores llamados clouders.

Actualmente la organización está mejorando distintos procesos, para que estos no sean un limitante en la escalabilidad, donde en su planificación está el de venta de rutas, actividad en la que se centra el trabajo. Para despachar los pedidos Wareclouds diariamente genera rutas las cuales contienen los pedidos, y cada clouder, cuando son publicadas, puede asignarse una al día para repartir los pedidos que contiene. La mejora de este proceso busca disminuir la cantidad de rutas que deben ser vendidas de forma manual, y así también aumentar la tasa de asignación de rutas que son tomadas automáticamente. Esta actividad, y la forma de realizarla, ha generado que existan rutas que no se vendan a tiempo, que genera pedidos retrasados, y su gestión dificulta la escalabilidad de Wareclouds, implicando incrementar próximamente de forma importante los trabajadores que deban dedicarse a esta actividad dada las proyecciones de crecimiento.

Dado que la promesa de valor es entregar los pedidos en menos de 24 horas en la Región Metropolitana, y para esto las 25 rutas diarias en promedio que son publicadas deben ser tomadas por los clouders, es de gran relevancia que el proceso funcione de forma adecuada para que se logre tal objetivo.

La forma en que se busca mejorar esta actividad consiste en utilizar herramientas de Machine Learning que permitan establecer las características más relevantes de los repartidores que determinan su reactivación, para así, mediante un modelo de regresión, conocer la propensión de reactivarse del grupo de clouders estudiado. Luego, clasificar los a los repartidores según su probabilidad de volver a tomar una ruta, y observar cómo varían los valores de los atributos más significativos. Como resultado principal se observa que los repartidores con mayor propensión en general optan por rutas de menor valor, se mantienen activos por más tiempo y toman rutas de forma periódica.

Finalmente, se plantea un prototipo que modifique aspectos en la mensajería y logre una disminución en el número de rutas que deban ser asignadas de forma manual, junto con corroborar los resultados para futuros proyectos de la organización. De esta forma el trabajo en su conjunto permitiría disminuir en un 60 % las personas necesarias para la venta de rutas, junto con un 65 % la fuga de marcas por incumplimiento en la promesa de valor.

Mamá,
por ti y para ti

Agradecimientos

Quiero comenzar agradeciendo a mi familia, a mis principales pilares, mi mamá Loreto y mi hermana Catalina, por estar en cada parte de este proceso, tanto en las penas como en las alegrías. También a la estrellita que tengo, que sé que de alguna forma estuvo en este duro camino.

También quiero agradecer a mis amigos, que han hecho estos años llevaderos, regalándome muy buenos recuerdos de la U. A somos 12, a mis amigas del colegio que estuvieron ahí siempre, a Felipe, Ignacio, a Mati.

Finalmente quiero agradecer a mi profesor guía, Juan Pablo Romero, que fue mucho más que eso. Gracias por estar, por el apoyo y la energía para sacar esto adelante.

Tabla de Contenido

1. Planteamiento del problema **1**

1.1. Antecedentes generales 1

1.1.1. Características de la organización 1

1.1.2. Actores relevantes 4

1.1.3. Desempeño organizacional 5

1.2. Problema y justificación 6

1.2.1. Área donde se desarrolla el trabajo 6

1.2.2. Problema identificado y su relevancia 7

1.2.3. Hipótesis y alternativas de solución 9

1.2.4. Justificación del problema 11

1.2.5. Valor generado a través de la solución 12

2. Objetivos **14**

2.1. Objetivo general 14

2.2. Objetivos específicos 14

2.3. Alcances 14

3. Marco conceptual **16**

3.1. Machine Learning 16

3.2. Entrenamiento y validación de modelos de Machine Learning 16

3.3. Técnicas para manejo de datos desbalanceados 18

3.3.1. Under-Sample 18

3.3.2. Over-Sample 18

3.3.3. SMOTE 19

3.4. Árboles de decisión (CART) 19

3.5. Regresión Logística 21

3.6. Métricas de desempeño de los algoritmos 22

3.6.1. Matriz de confusión 22

3.6.2. Curva ROC y AUC 24

4. Metodología **25**

5. Desarrollo metodológico **27**

5.1. Fase de comprensión del negocio 27

5.1.1. Entendiendo el problema u oportunidad 27

5.1.2. Los datos 29

5.2. Fase de recolección y comprensión de los datos 30

5.2.1. Base de datos 30

5.2.2. Atributos de la base 30

5.3. Fase de preparación de los datos 33

5.3.1. Limpieza de tablas 33

5.3.2. Consolidación de base de datos 34

5.3.3. Creación de atributos 36

5.3.4. Variable dependiente 41

5.3.5. Análisis descriptivo de los datos 42

5.4. Fase de modelado 47

5.4.1. Machine Learning para determinar las variables más importantes en la reactivación de los clouders 47

5.4.2. Machine Learning para conocer la propensión de reactivación de los clouders 49

5.5. Fase de evaluación 53

5.6. Fase de despliegue 55

5.6.1. Productivización del modelo 55

5.6.2. Impacto económico 56

5.6.3. Riesgos del proyecto 58

6. Conclusiones **59**

6.1. Cumplimiento de objetivos 59

6.2. Hipótesis planteadas 60

6.3. Impacto del trabajo realizado 61

6.4. Cambios en la metodología 62

Bibliografía **64**

Anexos **65**

Anexo A. **65**

Anexo B. **66**

Anexo C. **69**
C.1. Escenario optimista 69
C.2. Escenario medio 71
C.3. Escenario pesimista 73

Índice de Tablas

5.1. Tablas utilizadas y su respectivo número de instancias 31
5.2. Tabla Clouders y sus atributos 31
5.3. Tabla Clouders activos, Comunas, Comunas de rutas, Historial de rutas, Mensajería y Venta de rutas con sus respectivos atributos 32
5.4. Tablas consolidadas y sus atributos 36
5.5. Sectores y sus respectivas comunas 37
5.6. Sectores y sus respectivas comunas 38
5.7. Distribución de los sectores a los que pertenecen los clouders 39
5.8. Distribución de los sectores de preferencia de los clouders 39
5.9. Distribución de los sectores promocionados a los clouders 40
5.10. Distribución de la variable dependiente en la tabla Efecto mensajes 41
5.11. Distribución de la variable dependiente en la tabla Efecto clouders 41
5.12. Variables de los clouders que se consideran en el entrenamiento del modelo . 48
5.13. Desempeño del algoritmo Decision Tree 48
5.14. Variables más significativas que determinan la reactivación de un clouder . . 50
5.15. Desempeño del algoritmo Logistic Regression 50
5.16. Valor del error según punto de corte de propensión y peso de Error tipo I . . 54
5.17. Valor promedio de cada atributo según grupo de propensión 54

Índice de Ilustraciones

1.1. Organigrama de la empresa 2
1.2. Cantidad de pedidos mensuales 2021-2022 3
1.3. Número de clientes mensuales 2021-2022 3
1.4. Relación entre actores para el funcionamiento de la actividad 4
1.5. Ingresos mensuales de Wareclouds desde mayo a diciembre de 2021 5
1.6. Organigrama del equipo Operaciones 6
1.7. Vista de las rutas disponibles 8
1.8. Proceso de generación y venta de rutas 8
1.9. Pedidos mensuales y proyección de los próximos 6 meses 11
1.10. Rutas que deben ser vendidas mensualmente y proyección de estas los próximos 6 meses 12

3.1. Separación de datos en conjunto de entrenamiento y de prueba 17
3.2. Ejemplo de K-fold Cross-Validation (k=5) 18
3.3. Diagrama de árbol de decisión 20
3.4. Función logística 22
3.5. Matriz de confusión 22
3.6. Curva ROC 24

4.1. Diagrama metodología CRISP-DM 25

5.1. Mensaje de WhatsApp enviado a los clouders 29
5.2. Distribución de los sectores de preferencia de los clouders 42
5.3. Distribución de los sectores promocionados en los mensajes 42
5.4. Precio promedio de las rutas tomadas cuando los clouders estaban activos categorizados por sector de preferencia 43
5.5. Días activos de los clouders categorizados por sector de preferencia 44
5.6. Número de rutas tomadas por los clouders categorizados por sector de preferencia 44
5.7. Precio promocionado de los mensajes categorizados por el sector promocionado 45

5.8. Diferencia entre el precio promocionado en el mensaje y el precio promedio del clouder categorizado por el sector promocionado 45
5.9. Distribución de la variable "Contador de mensaje" 46
5.10. Curva ROC del algoritmo Decision Tree 49
5.11. Importancia de variables del algoritmo Decision Tree 49
5.12. Curva ROC modelo Logistic Regression 50
5.13. Curva de deciles y su respectiva propensión 51
5.14. Curva de deciles y su propensión según cada variable 52

A.1. Matriz de correlación de las variables de los clouders 65

B.1. Matriz de confusión de algoritmo Árbol de decisión con conjunto de entrenamiento 66
B.2. Matriz de confusión de algoritmo Regresión logística con conjunto de entrenamiento 67
B.3. Matriz de confusión de algoritmo Árbol de decisión con conjunto de test . . 67
B.4. Matriz de confusión de algoritmo Regresión logística con conjunto de test . . 68

C.1. Cálculo del ahorro en el costo de contratar nuevas personas dedicada a la venta de rutas manual 69
C.2. Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega 70
C.3. Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega en días de *peaks* de pedidos 70
C.4. Cálculo del ahorro en el costo de contratar nuevas personas dedicada a la venta de rutas manual 71
C.5. Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega 72
C.6. Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega en días de *peaks* de pedidos 72
C.7. Cálculo del ahorro en el costo de contratar nuevas personas dedicada a la venta de rutas manual 73
C.8. Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega 74
C.9. Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega en días de *peaks* de pedidos 74

Capítulo 1

Planteamiento del problema

1.1. Antecedentes generales

A continuación se desarrolla una presentación documentada de datos y antecedentes de Wareclouds, organización en la cual se realiza el trabajo, para entender el contexto en el cual se desenvuelve, junto con una descripción de los actores relevantes dentro de este.

1.1.1. Características de la organización

La empresa en la que se desarrollará el tema del proyecto es Wareclouds, la cual es una startup que entrega servicios de fulfillment y última milla a los ecommerce, fundada a fines de 2019 por Nicolás Aramayo y Arturo Quiroz. Nace con el objetivo de ofrecer puntos de retiro para productos vendidos de forma online por pequeñas y medianas empresas, pero debido a la pandemia y otros factores que dificultaban la escalabilidad de este modelo de negocio, termina convirtiéndose en una organización que ayuda a estas empresas a guardar, armar y despachar sus productos.

Este nuevo modelo, que considera el almacenamiento, armado y despacho de los productos comienza a mediados de 2020, y la diferencia respecto a otras organizaciones dedicadas al servicio de logística está en el cómo realizan tal actividad, ya que mediante un modelo colaborativo, cualquier particular puede implementar su casa como bodega o centro de distribución. Las personas que arman y almacenan los productos en sus casas son llamados wareclouds, mientras que quienes posteriormente realizan el despacho al consumidor final se denominan clouders.

Wareclouds tiene como misión optimizar y hacer eficiente la última milla, aspirando a ofrecer un servicio de entregas on-demand a pequeñas y medianas empresas, como también facilitar la venta de sus productos alrededor de todo el mundo. Hoy su propuesta de valor principal es entregar los pedidos dentro de la Región Metropolitana en menos de 24 horas, junto con ofrecer el precio más bajo de la competencia.

La organización actualmente cuenta con veinte trabajadores full-time, divididos en seis áreas. Uno de los fundadores se encuentra a cargo del área de Growth, dedicada principalmente a atraer nuevos clientes e inversionistas, junto con buscar estrategias para crecer continuamente. El segundo fundador, por su parte, lidera el área de Operaciones, la cual

como su nombre lo indica, se enfoca en las operaciones más importantes de la organización. El área de Customer Experience, liderada por la Head of Customer Experience, busca continuamente resolver cualquier problema que presenten principalmente tres de los actores de la actividad, los cuales son consumidores, clientes y despachadores. Tech es liderada por la Tech Lead y el Developer Senior, quienes buscan que el área trabaje en la mejora continua de las distintas plataformas desarrolladas por la organización. Finanzas, área que actualmente cuenta con un sólo trabajador, se enfoca en todas las obligaciones económicas de la empresa y de gestionar los ingresos y gastos que pueda tener. Finalmente, el área People & Culture, en la cual también se encuentra sólo un miembro de la organización, busca realizar lo necesario para generar un buen ambiente de trabajo y solucionar problemas de ámbito laboral y personal que se presenten.

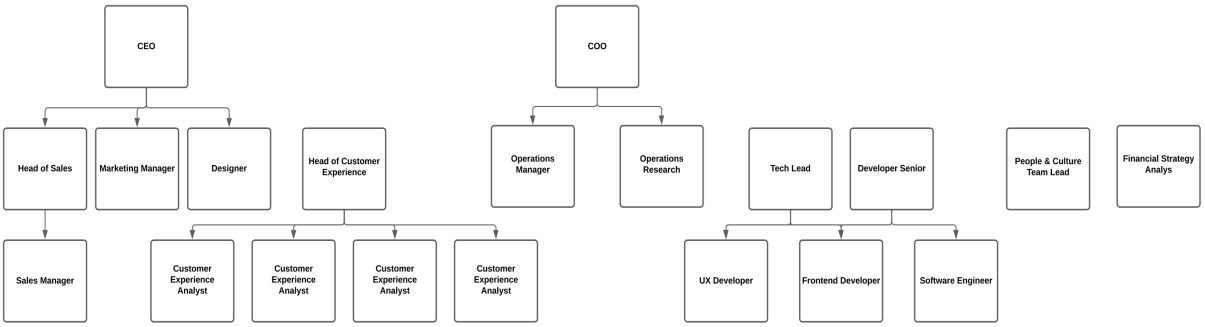

Figura 1.1: Organigrama de la empresa

Wareclouds ofrece tres servicios a los ecommerce. El primero es el almacenaje de los productos, el cual se cobra en base a la cantidad de pedidos mensuales, en un plan acorde a la etapa de la marca, existiendo cuatro rangos de número de pedidos: 0-10, 11-50, 51-200 y +200. Dichos productos son almacenados en la casa de los wareclouds, a quienes se les paga por el espacio utilizado por la marca. El segundo servicio corresponde al armado de pedidos, donde el cobro por armado es en base a la cantidad de SKUs, es decir, depende del número de productos distintos que ofrece el ecommerce. Similar al servicio de almacenaje, el warecloud recibe un monto fijo por cada pedido armado correctamente. Finalmente, el tercer servicio es el despacho de los productos, donde se define si el destino está dentro de la Provincia de Santiago, el resto de la Región Metropolitana u otras regiones. Si el despacho lo realiza un clouder, es decir, el destino se ubica en la Región Metropolitana, estos reciben un pago por la ruta que toman, la cual tiene un valor en función de la(s) comuna(s) que incluya y el número de pedidos.

Dado el modelo de negocios que tiene la empresa, sus ingresos están en función de los pedidos mensuales que debe despachar. Este número en gran parte del 2021 creció en promedio un 30 % al mes, mientras que los últimos meses tuvo una alta variabilidad. Sin embargo, de forma más indirecta, los ingresos de la empresa también dependen de la cantidad de clientes activos que tengan. Este número ha presentado desde enero de 2021 un crecimiento continuo, cerrando el año con cerca de 160 marcas. A continuación se observan los gráficos de los pedidos y clientes durante 2021-2022.

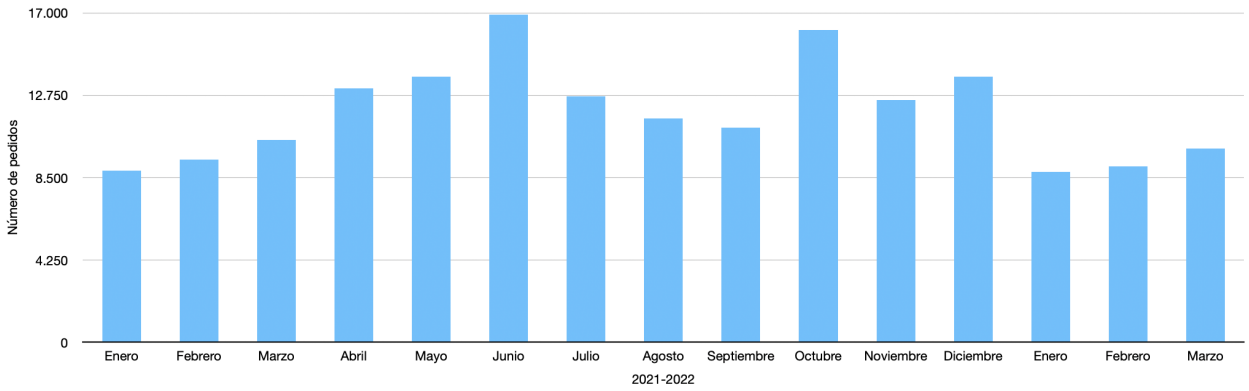

Figura 1.2: Cantidad de pedidos mensuales 2021-2022

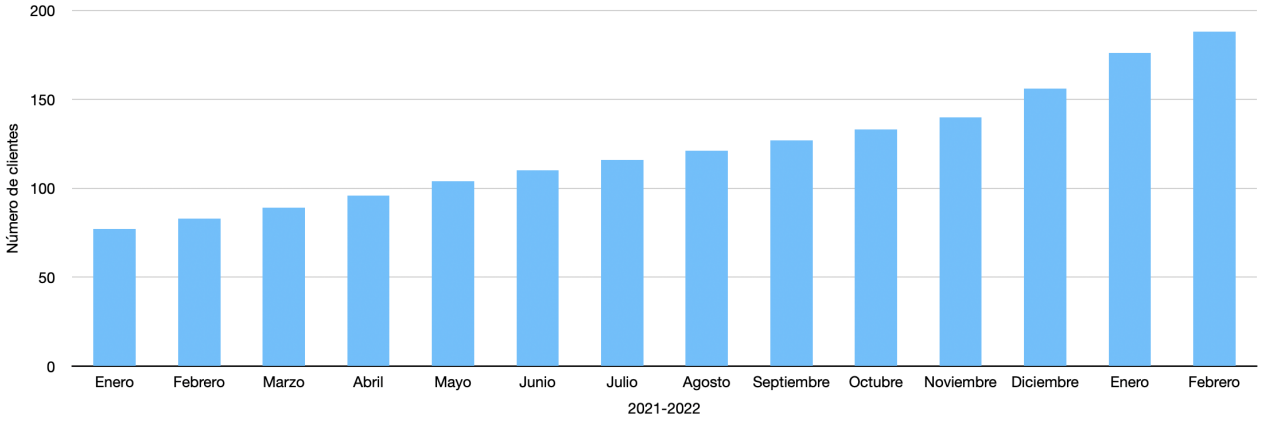

Figura 1.3: Número de clientes mensuales 2021-2022

Para llevar a cabo el servicio ofrecido a sus clientes, la empresa ha ido aumentando el número de wareclouds y clouders dependiendo del crecimiento que ha tenido. Hoy cuentan con 19 bodegas disponibles y 40 despachadores activos, los cuales han trabajado al menos una vez dentro de las últimas dos semanas en Wareclouds. La diferencia entre estos dos actores es la frecuencia con la que trabajan. Las bodegas deben estar disponibles de lunes a viernes para armar pedidos y entregarlos a los despachadores correspondientes, mientras que estos todas las mañanas, cuando se publican las rutas del día, tienen la posibilidad de elegir una si así lo desean, permitiéndoles trabajar los días que quieran, bajo una modalidad muy similar a Uber o Cabify.

Los competidores de Wareclouds se clasifican en tres tipos: Compañías de fulfillment tradicional, Dark Stores y empresas dedicadas a la última milla. Considerando lo anterior, sus principales ventajas competitivas son que presentan mayor flexibilidad y menor costo que las compañías tradicionales, debido a que no cuentan con recursos físicos, como también mayor rapidez en la entrega, permitiéndoles tener como propuesta de valor una entrega en la Región Metropolitana en menos de 24 horas, ya que están más cerca de la demanda al no tener bodegas en sectores periféricos. Es importante también mencionar que su modelo colaborativo es un punto importante para muchos clientes, ya que les otorga mayor cercanía con la logística de sus productos, pero sin estar haciéndose cargo de ella directamente.

1.1.2. Actores relevantes

Dentro de la actividad de la organización existen tres actores claves para esta. El primero es el ecommerce, quien es el cliente de Wareclouds y hace uso del servicio, los cuales buscan desligarse del problema de la logística, pero aún siendo parte de esta. Esto lo logra la marca ya que está en constante comunicación con su warecloud, donde este se encarga de todo lo que requiere el pedido del cliente final (consumidor), y la marca sólo debe ingresar la orden y los requerimientos que esta tenga. En la mayoría de los casos, las plataformas que los ecommerce utilizan para gestionar sus pedidos, como lo son Shopify, Woocommerce, Jumpselller, PrestaShop, entre otros, están integrados con las plataformas de Wareclouds, por lo que los pedidos que son ingresados en estas, caen directamente en las plataformas que tienen acceso los wareclouds y los ecommerce. El área de Customer Experience es la que se relaciona directamente con las marcas, siendo uno de los Analistas de este equipo el encargado de velar que estos reciban un buen servicio y responder frente a problemas que puedan surgir con sus productos y pedidos.

Un segundo actor es el warecloud, el cual está encargado de almacenar y armar los pedidos de las marcas que tiene asignadas. Los pedidos que ingresan en su plataforma antes de las 11 de la mañana, deben ser armados el mismo día, mientras que los siguientes deben estar armados antes del medio día del día siguiente, hora en la que comienzan a llegar los repartidores, y deben estar disponibles para entregarlos hasta las 16:00 hrs. Quien se relaciona con este grupo de personas es el Operations Manager de área de Operaciones, el cual está en constante comunicación con ellos, gestionando los quiebres de stock y los inconvenientes que estos puedan presentar con los pedidos.

Finalmente, un tercer actor es el clouder, el cual tiene como labor despachar los productos al consumidor. De lunes a viernes todos los repartidores que han trabajado alguna vez en Wareclouds reciben un mail a las 11:30 de la mañana con las rutas disponibles de ese día, las cuales tienen el número de pedidos que deben entregar y las direcciones de estos, junto con un valor asignado a esa ruta. Con esa información ellos deciden si la toman, donde si lo hacen, deben ingresar a la plataforma, seleccionar y tomar la ruta, y posteriormente, entregar todos los pedidos que esta contenga el mismo día. Quien se preocupa de promocionar las rutas y se comunica con ellos si existe algún problema durante el trayecto es uno de los Customer Experience Analyst. A continuación, se observa cómo se relaciona cada actor para este servicio de logística.

Figura 1.4: Relación entre actores para el funcionamiento de la actividad

1.1.3. Desempeño organizacional

Wareclouds se encuentra en una importante etapa de crecimiento. Como se mencionó previamente, esto se observa en el número de pedidos y marcas que tiene la empresa, los cuales tienen directa relación con los ingresos mensuales que genera. Este número de pedidos también son un indicador importante para sus planes a futuro, los cuales se enfocan en abrir operaciones en México, para posteriormente llegar a Estados Unidos. A continuación se observa un gráfico que detalla los ingresos de los últimos meses de la organización.

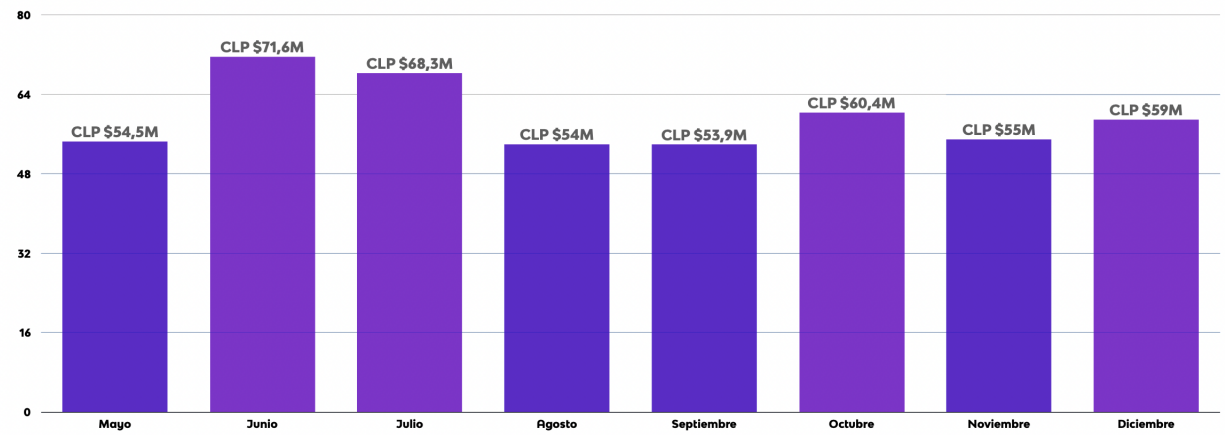

Figura 1.5: Ingresos mensuales de Wareclouds desde mayo a diciembre de 2021

Existen tres principales hitos importantes que busca completar la compañía dentro de este año y que se traducirían en un crecimiento acelerado de Wareclouds. El primero corresponde a contar, desde el 1 de abril, con un nuevo servicio que permita entregar en menos de 90 minutos los pedidos que así sean seleccionados por los consumidores. Este servicio busca seguir transformando la logística en Chile, dado que sería posible ofrecerlo con precios muy inferiores a los de la competencia, debido al modelo de negocios que tiene la empresa. Este nuevo servicio, según las proyecciones, implicaría empezar a aumentar los pedidos y clientes de forma considerable.

De la mano con lo anterior, Wareclouds recientemente cerró una ronda de inversión de $1,4 millones de dólares, capital con el que espera comenzar a operar en México los próximos meses y triplicar su operación, considerando que en ese país hay alrededor de cuatro millones de Pymes. Estas empresas se estima que generan el 72 % del empleo y 52 % del PIB, pero que sin embargo presentan problemas de transformación digital y logísticos, puntos en los que Wareclouds puede ser clave.

Finalmente, su próximo paso, a fines de este año si el crecimiento se alinea con las proyecciones previas, es entrar en el mercado Estadounidense, lo cuál está condicionado a tres variables que son: tener más de 100.000 pedidos mensuales, una tasa de servicio del 97 % y un NPS (Net Promoter Score) superior a 50, el cual es un indicador de la experiencia del cliente y mide las probabilidades de que estos recomienden la empresa.

1.2. Problema y justificación

A continuación se realiza una descripción del contexto del problema que se aborda, junto con la justificación y valor agregado de llevarlo a cabo.

1.2.1. Área donde se desarrolla el trabajo

La investigación se desarrolla en el área de Operaciones de Wareclouds, la cual está conformada por tres personas, donde cada una desempeña distintas funciones, pero que dada las características de sus responsabilidades, se considera adecuado subdividirla en dos. La primera compuesta por el Operations Manager, que tiene como objetivo ser un canal directo entre la organización y las bodegas. Mientras que la otra parte del área, donde se encuentra el Chief Operating Officer y el Operations Research, busca continuamente mejorar la operación interna, mediante estudios y manejo de datos, junto con preocuparse de que los procesos no dificulten la escalabilidad de la organización, sino que por el contrario, permitan potenciar su crecimiento. A continuación se observa el organigrama del área.

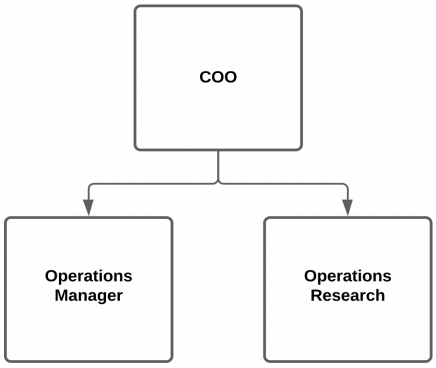

Figura 1.6: Organigrama del equipo Operaciones

Como ya se mencionó, una parte de Operaciones se relaciona directamente con las bodegas, donde este trabajador está encargado de todo lo relacionado con los wareclouds, y sus principales funciones consisten en enviarles stock de los productos que manejan cuando lo solicitan, atender cualquier problemática que presenten con las marcas y/o con las plataformas, y desarrollar manuales que les faciliten el uso de nuevas funcionalidades de la plataforma.

Ambos integrantes de la otra parte del área diariamente desempeñan funciones relacionadas con la optimización de procesos y manejo de datos que permita escalar la operación mediante distintas herramientas computacionales. Actualmente, el COO se encuentra principalmente trabajando en la implementación de un nuevo servicio que permita entregar los pedidos de la Región Metropolitana en 90 minutos, y el Operations Research está enfocado en la automatización de devoluciones de pedidos que son despachados por Blue Express, es decir, los pedidos que son devueltos desde otras regiones. Este último proyecto busca implementar un servicio que permita que estos productos lleguen nuevamente a los wareclouds asociados al ecommerce que los vendió.

Considerando las tareas que son realizadas por esta parte del área, y dado que uno de sus KPIs corresponde a la tasa de pedidos entregados a tiempo, dentro de su planificación se

encuentra un proyecto que busca mejorar la asignación de las rutas que no son seleccionadas automáticamente, debido a que la forma en que esta se aborda, no es factible a medida que los pedidos aumentan, y dificulta el cumplimiento de la entrega en menos de 24 horas. Esta mejora se traduciría directamente en un aumento del número de rutas diarias vendidas y disminución de los pedidos retrasados, aportando a la escalabilidad de Wareclouds y no siendo un impedimento en el crecimiento de la organización. Debido al problema que esta actividad genera, se cree atingente que se desarrolle el trabajo de título dentro de esta parte del área, con el objetivo de modificar este proceso de forma que se alinee con el desarrollo de la empresa.

Si bien el trabajo se desarrolla principalmente en el área de Operaciones, en base a lo detallado anteriormente, quien hoy desempeña la actividad crítica del proceso que se desea abordar pertenece al equipo de Customer Experience. Uno de estos trabajadores, quien se desempeña como Customer Experience Analyst, es el que diariamente debe comunicarse con distintos clouders para promocionarles las rutas que quedan disponibles, luego de estas ser publicadas y no ser seleccionadas por ningún repartidor naturalmente. También, dentro de sus labores diarias se encuentran atender cualquier problema que los clouders presenten durante la entrega de pedidos, siendo los más recurrentes no encontrar al consumidor en la vivienda y presentar dificultades con la plataforma, junto con hacerle la capacitación correspondiente a los nuevos clouders. El organigrama de esta área se presenta a continuación.

1.2.2. Problema identificado y su relevancia

El problema que hoy se presenta en la organización está relacionado con el proceso de venta de rutas diarias, a través de las cuales se entregan los pedidos, en días de alta demanda y a medida que el número de pedidos aumenta. Este viene precedido por la generación de rutas, actividad que se realiza de lunes a viernes, y donde se seleccionan todos los pedidos que han sido ingresados el día previo después de las 11 de la mañana y antes de esa misma hora de ese día, junto con los pedidos pendientes, los cuales corresponden a pedidos que por distintas razones no fueron entregados los días previos. Es decir, las rutas generadas el día martes, por ejemplo, contienen los pedidos realizados por los consumidores el día lunes de esa semana, después de las 11 de la mañana, hasta los hechos el martes antes de esa misma hora. Para el caso de los lunes, se consideran los pedidos que han ingresado desde el viernes, posterior a las 11:00 hrs. y previo a esa misma del día lunes, incluyendo los que cayeron sábado y domingo. Los pedidos seleccionados se asignan mediante un modelo realizado por Wareclouds, a determinadas rutas, en función de las comunas en las que deben ser despachados. Posteriormente, estas rutas son publicadas, alrededor de las 11:30 hrs. en la plataforma de los clouders, como se muestra a continuación.

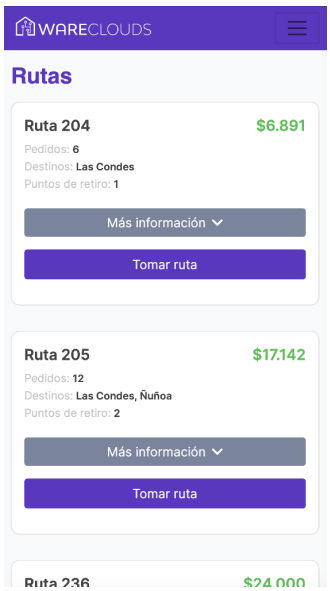

Figura 1.7: Vista de las rutas disponibles

Luego de ser publicadas las rutas de ese día, se les envía un mail a todos los repartidores que han trabajado alguna vez en Wareclouds, informándoles que hay rutas disponibles, para que ellos posteriormente elijan la que más les acomoda tomar.

La mayoría de las rutas son rápidamente seleccionadas por los clouders, sin embargo, existen otras que suelen quedar pendientes y que deben intentar ser vendidas de forma manual. A continuación, se observa el proceso de generación y venta de rutas.

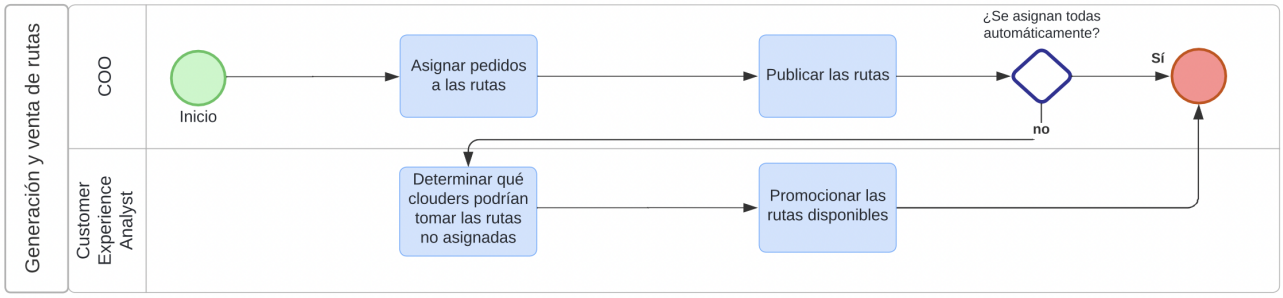

Figura 1.8: Proceso de generación y venta de rutas

En términos de cifras, hoy se entregan cerca de 10.000 pedidos mensuales, lo que corresponde en promedio a 500 pedidos diarios que deben ser despachados durante el mismo día. Dado que cada ruta tiene asignado un número cercano a los 25 pedidos, se estima que son 20 rutas las generadas diariamente.

Respecto a los repartidores, son 40 los clouders activos, es decir, que han tomado al menos una ruta durante las últimas dos semanas. Mientras que son 440 los clouders que componen la base de datos, es decir, que han repartido al menos una vez en Wareclouds.

Con esta cantidad de pedidos y repartidores disponibles, el 80 % de las rutas son tomadas automáticamente por los clouders, sin embargo, en un día de 800 pedidos, esta tasa suele bajar hasta el 60 %. Esto implica que entre un 20 % y 40 % de las rutas diarias no se asignan

naturalmente a ningún repartidor, debiendo existir una segunda instancia en la que se deban gestionar de forma manual.

Las rutas que no son asignadas automáticamente se intentan promocionar de forma manual. Es aquí donde uno de los Customer Experience Analyst debe intentar determinar cuáles son los clouders con más probabilidades de tomar las rutas que quedan disponibles, dado el historial de ruta(s) previamente tomada(s) por estas personas, y posteriormente comunicarse con ellos para ofrecerles una similar, que el trabajador cree que podría interesarle. La decisión del trabajador para ofrecer estas rutas se realiza entonces sólo en base al juicio de la persona y dada la capacidad que esta tenga de ver qué clouders han tomado antes rutas similares y/o han trabajado en Wareclouds durante el último tiempo.

A la escala actual, son en promedio 100 los pedidos mensuales que se retrasan debido a que estaban asignados a una ruta que no se pudo vender a tiempo, lo que corresponde a 4 rutas al mes (5 % de las rutas mensuales que deben ser vendidas manualmente). Esta situación vislumbra los primeros problemas que se presentan en este proceso, y es que no todas las rutas se logran vender a tiempo, lo que por consecuencia tiene que no todos los pedidos cumplen la promesa principal de la empresa que corresponde a entregas en menos de 24 horas.

1.2.3. Hipótesis y alternativas de solución

Como se detalló previamente, el problema que se aborda dentro de este trabajo de título corresponde a que la totalidad de las rutas diarias publicadas por Wareclouds no se vende, es decir, no se generan todos los matches necesarios entre estas y los clouders. Esto afecta directamente a todos los clientes y a la reputación de la empresa, dado que todos los pedidos de las rutas que quedan sin asignar son retrasados y por lo que no se cumple la promesa de valor que consiste en despachar los pedidos en menos de 24 horas. Junto con esto, el tiempo que implica vender las rutas de forma manual, dificulta en gran medida la escalabilidad de la operación, ya que considerando la proyección de pedidos que tiene Wareclouds para los próximos meses, se debería aumentar constantemente el número de trabajadores dedicados a esta actividad.

Dentro de las hipótesis respecto a las causas que generan este problema en la venta de rutas, un punto importante es que los pedidos se concentran en el sector oriente de la Región Metropolitana. Esto ha implicado que los wareclouds se ubiquen en comunas como Las Condes, Providencia y Ñuñoa, junto con que la mayoría de las rutas tenga como destinos estas comunas.

Dado que los pedidos deben ser primero retirados por los clouders en las bodegas correspondientes, esto genera que exista una mayor distancia entre los puntos de retiro y de entrega cuando las rutas van a comunas que no pertenecen al sector oriente, y por lo tanto, sean menos atractivas para la mayoría de los repartidores. Sin embargo, existen personas que si se ven interesadas en rutas que se dirigen hacia otros sectores, principalmente debido a que sus precios son mayores y/o viven en esas comunas.

Por otra parte, ya que cerca del 80 % de las rutas contienen pedidos cercanos a los puntos de retiro, el otro 20 % presenta una gran variabilidad de sus destinos. De forma más concreta, el número de rutas mensuales que se dirige a una determinada comuna, fuera del sector oriente, es mucho menor que las que se encuentran dentro de ese sector, estando todos los

días disponibles, a diferencia del 20 % restante.

Debido a esta desproporción en la frecuencia de los destinos de las rutas, se cree que los clouders interesados en comunas con menos concentración de pedidos, con el tiempo dejan de observar las rutas diarias disponibles dado que son pocas las veces que observan rutas de su interés. Es por esto, que se considera atingente y de gran valor, tanto para la organización como para los repartidores, promocionarles de forma rápida y directa las rutas disponibles acordes a su preferencia. El beneficio para Wareclouds es disminuir el tiempo en encontrar al clouder correcto y lograr que este tome la ruta que está disponible, mientras que para este repartidor sería beneficioso saber cuando son publicadas las rutas de su interés.

Otra hipótesis que se plantea respecto a las causas que contribuirían a la falta de toma de rutas, se relaciona con la forma de comunicación e incentivos que se tienen hoy. En relación al tipo de comunicación, se cree que un correo electrónico no siempre es instantáneo, dado que muchas veces se presentan variaciones de tiempo en las notificaciones. Esto podría estar generando que los clouders que no están continuamente revisando su correo, sean notificados con las rutas disponibles en distintos momentos, lo que finalmente provoca un sesgo no deseable en las rutas que cada repartidor ve.

Respecto a los incentivos, si bien hoy Wareclouds presenta algunos, estos están enfocados principalmente en la fidelización del clouder, y los otorga en función del número de rutas que los repartidores tomen a la quincena y el rating con el que sean evaluados por los wareclouds. Se cree que esto podría no estar alineado con los intereses de determinados repartidores, que buscan trabajar de forma esporádica y no estar constantemente trabajando en la organización.

En vista de lo anterior, se concluye principalmente que las causas del problema se relacionan por una parte, con que no todos los clouders desean tomar todas las rutas, sino que las rutas que seleccionan están en función de sus intereses personales y lugar donde reside. Por otra parte, debido a la heterogeneidad de los destinos de las rutas, esto genera que los repartidores interesados en las rutas con destinos menos frecuentes, pierdan interés en Wareclouds y opten por no revisar la plataforma continuamente, disminuyendo la probabilidad de que estas sean tomadas. Finalmente, también se considera la posibilidad de que las formas de comunicarse con los clouders e incentivarlos a tomar rutas podría no estar siendo la óptima, sobre todo para repartidores que desean trabajar en Wareclouds pero no de forma tan continua.

Teniendo en cuenta las causas mencionadas, existen diversas soluciones que podrían presentarse para afrontar el problema presente. Una de ellas corresponde a automatizar la venta de rutas, con el objetivo de promocionarlas de forma personalizada a los repartidores, en función de sus intereses personales y el lugar donde residen. Esta solución requeriría gran cantidad de datos para conocer suficiente a los clouders, teniendo un historial considerable de las rutas que se asignaron, junto con otras características importantes que puedan permitir clasificarlos de forma correcta y promocionarles rutas que se alineen con sus intereses.

Otra posible solución sería establecer una forma de comunicación más directa e instantánea con los repartidores que tienen interés en las rutas con destinos menos frecuentes, para así informarles cuando estás estén disponibles sin necesidad de que ellos deban estar constantemente revisando la plataforma. De forma concreta sería promocionarles las rutas disponibles que podrían ser de su interés mediante Whatsapp. Si bien aumentaría las vistas de las rutas, esto podría incurrir en altos costos para la empresa dado que la mensajería mediante esta plataforma tiene un alto valor y el objetivo de utilizarla es poder focalizar su

uso lo más posible.

Una tercera alternativa de solución está relacionada con cambios en los canales de comunicación e incentivos que se le ofrecen a los repartidores, buscando aumentar el número de reactivaciones en clouders que podrían tener interés en trabajar de forma menos continua. Esta alternativa busca ser un poco menos riesgosa que las mencionadas anteriormente, pudiendo así combinar tipos de mensajería e incentivos, dependiendo de los clouders a los que se esté notificando, sin utilizar una única solución para todos los repartidores.

Para las alternativas de solución presentadas se requiere identificar previamente los intereses y características más relevantes de los repartidores que se han intentado reactivar mediante mensajería, para posteriormente automatizar estos mensajes, optar por otro tipo de comunicación y/o proponer nuevos incentivos. Es por esto que en este trabajo se busca determinar esos atributos más relevantes de los clouders, para posteriormente proponer un piloto experimental que varíe tipos de comunicación e incentivos en busca de aumentar las reactivaciones, así como también, corroborar el comportamiento de los clouders según su propensión de volver a trabajar en Wareclouds, pudiendo posteriormente llevar a cabo otra de las soluciones.

Dado que la solución busca aumentar el conocimiento de los repartidores que se inactivan, para así proponer cambios en la mensajería con foco en aumentar las reactivaciones de los clouders, este proyecto se alinea con modificar un proceso para que este no sea un limitante en la escalabilidad de Wareclouds y permita potenciar su crecimiento. Este objetivo tiene un enfoque de procesos, algo que caracteriza a un egresado de Ingeniería Civil Industrial, junto con abordar un problema de gestión, en busca de una mejor solución que permita obtener resultados importantes para el crecimiento y continuidad de la organización.

1.2.4. Justificación del problema

Teniendo en cuenta la etapa en la que se encuentra Wareclouds, donde tiene planificado comenzar a operar en México en Mayo de 2022, y en Estados Unidos a fines de este año, se espera que el número de pedidos mensuales crezca de forma acelerada, observándose la proyección a continuación.

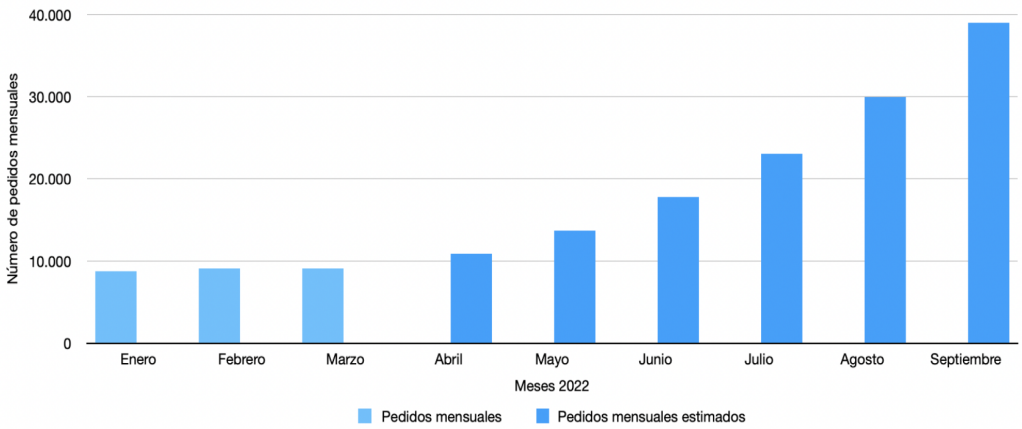

Figura 1.9: Pedidos mensuales y proyección de los próximos 6 meses

Como se observa, se espera que a fines de julio la operación se duplique y en 6 meses más los pedidos se cuadrupliquen. Frente a este pronóstico acelerado de crecimiento en las ventas, esto se traduciría también en un aumento importante de las rutas que deban ser vendidas manualmente.

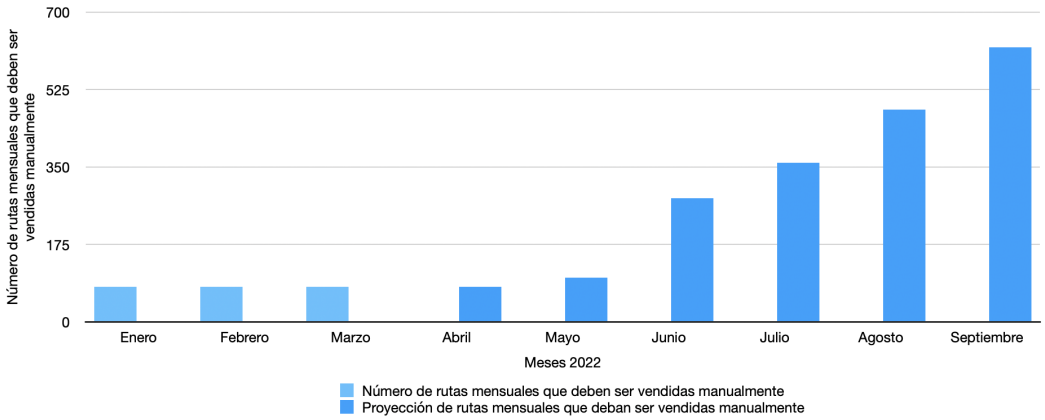

Figura 1.10: Rutas que deben ser vendidas mensualmente y proyección de estas los próximos 6 meses

En vista de lo anterior, a fines de julio se requerirían 5 personas para realizar la venta de rutas manual, es decir, duplicar la operación implicaría agregar a 4 trabajadores que realicen esta tarea. En septiembre, donde Wareclouds espera estar despachando en promedio 40.000 pedidos mensuales, serían cerca de 620 rutas las que deban ser vendidas de forma manual durante ese mes, lo que significaría tener a 8 trabajadores dedicados a esa función.

Junto con lo anterior, a la escala actual hoy Wareclouds presenta una fuga cercana al 2 % de sus clientes debido al incumplimiento en los plazos de entrega de los pedidos, lo que se traduce en promedio en tres marcas al mes. Teniendo presente el crecimiento que tendría en los próximos meses la empresa, a fines de septiembre serían cerca de 130 marcas las que decidirían cancelar el servicio por esta razón.

Finalmente, hoy también existen dificultades de cumplir con la promesa de valor en fechas con *peaks* de pedidos, como los son festividades o Cyber Days. Aquí los pedidos suelen aumentar en promedio un 60 %, lo que dificulta aún más las entregas el mismo día. A la escala actual, en cada evento donde los pedidos aumentan, son en promedio tres las marcas que cancelan el servicio, y donde estas suelen ser ecommerce grandes, generando un costo mayor que las marcas promedio en la empresa.

El considerable crecimiento del número de trabajadores necesarios para gestionar la venta de rutas manual a medida que aumentan los pedidos mensuales, junto con el aumento de la fuga de clientes a medida que la empresa crece, dificulta la escalabilidad de la organización, por lo que se considera que el proceso actual hace inviable soportar ese crecimiento y entregar los pedidos dentro del plazo acordado.

1.2.5. Valor generado a través de la solución

Teniendo en cuenta el problema que se ha detallado previamente, el valor principal de este proyecto es aumentar el conocimiento de los clouders, uno de los tres actores más rele-

vantes de la organización, así como también, plantear una propuesta experimental, en base a lo estudiado, que permita mejorar el cumplimiento de la promesa de valor que tiene la organización en la Región Metropolitana.

Con el número de pedidos mensuales que hoy presenta la organización, son cerca de 100 los pedidos que se retrasan al mes debido a que las rutas en las que estos estaban asignados no se lograron vender a tiempo. Esto implica que a esta escala, el 5 % de las rutas mensuales que deben ser vendidas de forma manual, no logran asignarse a un repartidor.

Considerando el acelerado crecimiento que espera tener Wareclouds dentro de los próximos 6 meses, donde al final de este período esperan cuadruplicar sus ventas, mantener el proceso que se lleva a cabo sin ninguna intervención de mejora, implicaría aumentar de forma poco escalable los trabajadores dedicados a esta función, junto con presentar un aumento importante del número de marcas fugadas debido al incumplimiento de la promesa de valor.

La primera parte del proyecto está enfocada en generar valor mediante el aumento del conocimiento de los clouders que han dejado de trabajar en Wareclouds, ya que que no existe ningún proyecto previo relacionado con esto, pudiendo así obtener primeros *insights* respecto a los repartidores. La segunda parte del trabajo, se enfoca entonces en plantear una propuesta experimental, basada en los resultados obtenidos en la primera parte, que disminuya el número de rutas que deban ser vendidas de forma manual mediante el aumento de reactivaciones de los clouders. Con esto, a la vez se espera tener un proceso que sea más escalable, junto con disminuir la fuga de clientes debido al incumplimiento en los plazos de entrega.

Capítulo 2

Objetivos

A continuación se define el objetivo general, los objetivos específicos y los alcances del trabajo.

2.1. Objetivo general

El objeto de este trabajo es diseñar un proyecto que aumente el conocimiento de los repartidores para mejorar la actividad de reactivación de estos y así aumentar la asignación de rutas dentro del mismo día y cumplir el compromiso de servicio, a partir del uso de datos y algoritmos de Machine Learning en una empresa de logística.

2.2. Objetivos específicos

1. Comprender el negocio y la relevancia de los clouders en este.
2. Identificar las instancias donde los clouders vuelven a tomar una ruta luego de recibir un mensaje.
3. Determinar las variables más importantes que influyen en la reactivación de un clouder.
4. Clasificar a los clouders según su propensión de volver a tomar una ruta.
5. Proponer un diseño experimental modificando variables en los mensajes para validar la hipótesis de propensión.

2.3. Alcances

En términos simples, el proyecto busca determinar qué características de los clouders son más relevantes en la reactivación de un repartidor, para así focalizar el envío de los mensajes. Dado que para esto se utiliza Machine Learning, disciplina en la cual mientras más datos existen, mayor es el aprendizaje de los algoritmos, los datos utilizados son los recolectados hasta el 24 de mayo de 2022, y no se considera ningún mensaje enviado posterior a esa fecha. Esto con el fin de cumplir el plazo establecido para el término del proyecto.

Un segundo alcance relevante en el proyecto, es que dado que este busca ser el comienzo de una posterior automatización de venta de rutas, el proyecto finaliza con la propuesta de un diseño experimental que permita validar la hipótesis relacionada con las características de los clouders y su respectiva propensión para volver a trabajar en Wareclouds, pero no se considera como parte de este la automatización.

Capítulo 3

Marco conceptual

A continuación se describe el marco conceptual que tiene el propósito de dar al proyecto un sistema coordinado y coherente de conceptos y proposiciones que permitan abordar el problema de forma correcta.

3.1. Machine Learning

El Aprendizaje Automático o Machine Learning (ML) corresponde a una de las ramas de la Inteligencia Artificial (IA) que le permite a un sistema aprender de los datos, en lugar de hacerlo mediante la programación explícita. Esta rama utiliza una variedad de algoritmos que aprenden iterativamente de los datos para así mejorar la descripción y predicción de los resultados.

Un modelo de Aprendizaje Automático es la salida de información que se genera cuando se entrena un algoritmo con datos. Luego del entrenamiento, al proporcionar un modelo con una entrada, se le dará una salida. Dependiendo de la naturaleza del problema, existen diferentes enfoques basados en el tipo y el volumen de datos. En este proyecto en particular, luego de la comprensión y preparación de los datos, se utilizan dos algoritmos para obtener las variables más importantes que determinan el comportamiento de los actores, y para clasificarlos según su probabilidad de tener un determinado comportamiento.

3.2. Entrenamiento y validación de modelos de Machine Learning

Los modelos de Machine Learning aprenden de los datos con los que son entrenados, y a partir de ellos intentan encontrar o inferir el patrón que les permita predecir el resultado para un nuevo caso. Sin embargo, para poder calibrar y evaluar si un modelo funciona, se necesita probarlo con un conjunto de datos diferente. Por ello, en todo el proceso de aprendizaje, los datos de trabajo se dividen en dos partes: datos de entrenamiento (training) y datos de prueba (test).

- **Conjunto de entrenamiento (training)**: Datos con los que se entrena el modelo.

- **Conjunto de prueba (test)**: Datos que el modelo no ha visto y que se utiliza para ajustar parámetros y seleccionar el mejor algoritmo.
- **Error de entrenamiento**: Error que comete el modelo al predecir observaciones que pertenecen al conjunto de entrenamiento.
- **Error de test**: Error que comete el modelo al predecir observaciones que pertenecen al conjunto de prueba.

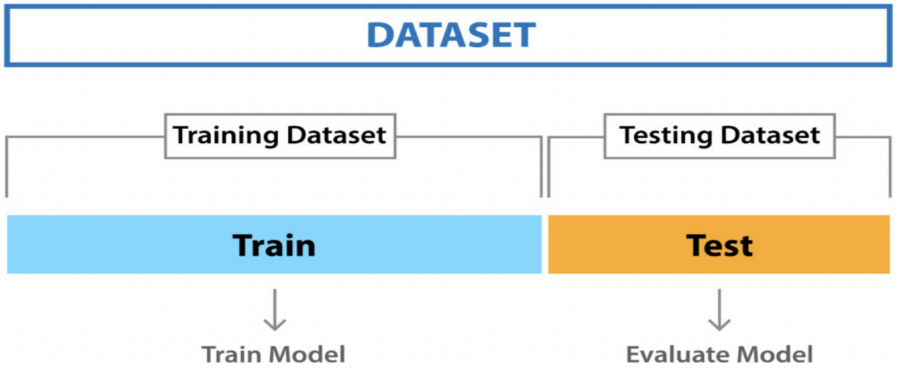

Figura 3.1: Separación de datos en conjunto de entrenamiento y de prueba

De esta forma, se tiene entonces un set de entrenamiento y un set de datos de validación. Esta corresponde a la forma más simple de entrenar y validar modelos, por lo que puede surgir un inconveniente cuando se dispone de una cantidad limitada de datos. Esto debido a que la cantidad de conjuntos de entrenamiento pueden ser tan pequeños que no se consiga un modelo efectivo. Para hacer frente a este problema, existen estrategias de validación, en donde a continuación se detallan las más utilizadas.

1. **Validación simple**: Esta técnica consiste en dividir aleatoriamente las observaciones disponibles en dos grupos, uno para entrenar y otro para evaluar. Los problemas que presenta esta estrategia son:
    - La estimación del error es variable dependiendo de las observaciones que se incluyan en cada grupo.
    - Se pierde poder predictivo, ya que se dispone de menos información para entrenar el modelo.
2. **K-fold Cross-Validation**: Esta estrategia consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño y k-1 grupos se utilizan para entrenar el modelo, mientras que uno se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración. Así es como el proceso genera k estimaciones del error y donde el promedio se utiliza como estimación final.

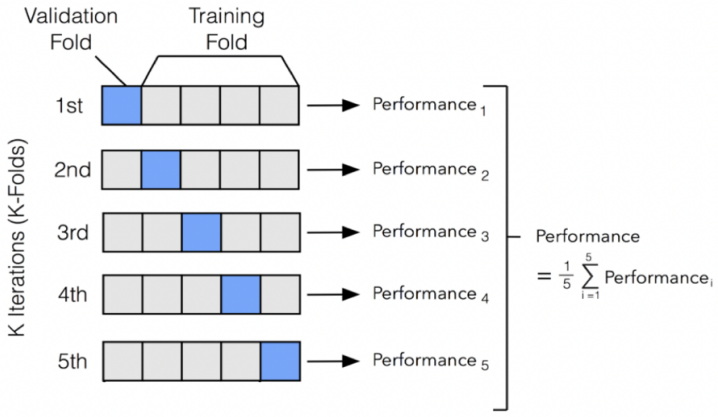

Figura 3.2: Ejemplo de K-fold Cross-Validation (k=5)

3. **Repeated K-fold Cross-Validation**: Esta última estrategia es similar a la descrita previamente pero repitiéndolo n veces.

3.3. Técnicas para manejo de datos desbalanceados

Los conjuntos de datos desbalanceados son bastante comunes en todas las industrias, en este caso, los mensajes que no tienen efecto sobre los clouders son más que los que si logran una reactivación, así como también, los repartidores que no vuelven a tomar una ruta luego de inactivarse son más que los que si lo hacen.

Para solucionar esto existen diferentes métodos, el más sencillo consiste en muestrear aleatoriamente el conjunto de datos de entrenamiento, ya sea, eliminando registros de la clase mayoritaria (random undersampling) o duplicando ejemplos de la clase minoritaria (random oversampling). Al usar undersampling existe el inconveniente de eliminar registros de la clase mayoritaria, lo que puede resultar en la pérdida de representatividad de la muestra y generar sesgo en el modelo. Al usar oversampling existe el riesgo de sobreajuste debido a que se duplica información existente. Una forma de reducir estos riesgos es utilizando ambos enfoques en menor proporción sobre la muestra.

3.3.1. Under-Sample

Este primer método se basa en balancear la distribución de los datos eliminando instancias de la clase mayoritaria. A pesar de su sencillez y de la posible reducción de tiempo de procesamiento de datos, presenta el riesgo de eliminar elementos de la muestra potencialmente importantes, por lo que se han desarrollado otros capaces de realizar una selección más inteligente sobre los elementos del conjunto de datos de la clase mayoritaria.[1].

3.3.2. Over-Sample

De forma contraria al método previo, esta técnica consiste en seleccionar y duplicar de forma aleatoria datos de la clase minoritaria hasta que esta coincida con el número de la

clase con más observaciones. En este caso existe el riesgo de sobre ajustar el modelo.[1].

3.3.3. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) es una técnica estadística que busca aumentar el número de casos de la clases con menos observaciones, pero a diferencia del Over-Sample, las nuevas instancias se crean artificialmente a partir de un algoritmo que utiliza las características disponibles para cada clase, y hace que las muestras sean más generales. En este método primero se selecciona aleatoriamente un registro de la clases minoritaria y se encuentran los k vecinos de clase minoritaria más cercanos. Posteriormente, el registro sintético se crea eligiendo aleatoriamente uno de los k vecinos más cercanos y conectando ambos para formar un segmento de línea en el espacio de características. Las instancias sintéticas se generan como una combinación convexa de las dos instancias elegidas.[2].

3.4. Árboles de decisión (CART)

Son un tipo de algoritmo supervisado desarrollado por Breiman, Friedman, Olshen y Stone en 1984. Dependiendo del tipo de variable a predecir, se diferencian dos tipo de árboles, los de clasificación y los de regresión. Los de clasificación tienen una variable a predecir de tipo categórica y el valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que han "caído" en esa región. Por otra parte, los de regresión tienen una variable numérica a predecir, y los valores de los nodos terminales se reducen a la media de las observaciones en esa región [14].

En el desarrollo del proyecto se opta por utilizar este algoritmo para determinar cuáles son las variables más relevantes en las reactivación de los clouders, dado que dentro de sus ventajas está el permitir identificar de forma rápida y eficiente los predictores más importantes.

Terminología

- **Nodo raíz**: Representa a toda la población o muestra, y esta se divide en dos o más conjuntos homogéneos.
- **Nodos de decisión**: Corresponde a un sub-nodo cuando este se divide en sub-nodos adicionales.
- **Hojas o nodos terminales**: Son llamados así los nodos que no tienen una división adicional (sin hijos).
- **Ramificación o división**: Es un proceso de división de un nodo en dos o más sub-nodos.
- **Rama o sub-árbol**: Corresponde a una sub-sección del árbol de ddecisión.
- **Poda**: Es el proceso donde se reduce el tamaño de los árboles de decisión eliminando determinados nodos.
- **Nodo padre-hijo**: El nodo que se divide en sub-nodos se denomina nodo principal, mientras que los sub-nodos son hijos del nodo principal.

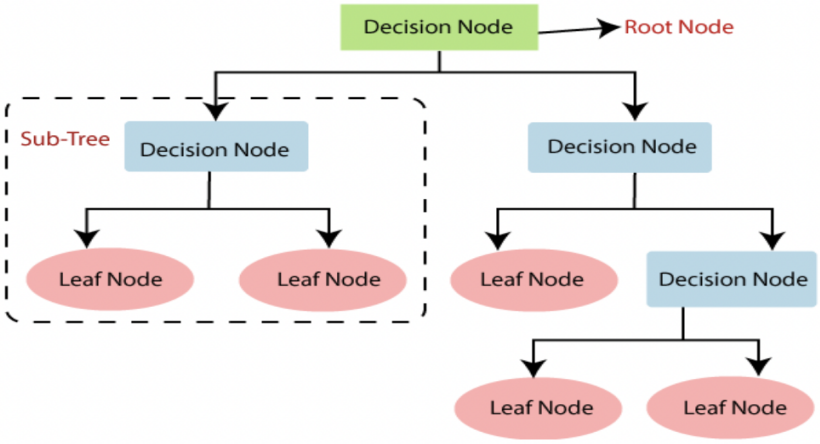

Figura 3.3: Diagrama de árbol de decisión

Algoritmo del árbol de decisión

Los árboles de decisión clasifican las observaciones ordenándolas en el árbol desde la raíz hasta un nodo hoja/terminal, y el nodo hoja/terminal proporciona la clasificación de la observación. Cada nodo del árbol actúa como un caso de prueba para algún atributo, y cada borde que desciende del nodo corresponde a posibles respuestas al caso de prueba. Este proceso es de naturaleza recursiva y se repite para cada subárbol que desciende del nuevo nodo. La técnica del árbol de clasificación sigue los siguientes pasos:

1. Comienza en el nodo raíz, que contiene el conjunto de datos completo.
2. En cada nodo se toma la decisión de ramificar o detenerse.
3. Cuando se toma la decisión de parar, se elige la etiqueta del nuevo nodo hoja.
4. Cuando se toma la decisión de ramificar, se elige qué variable se usará para la ramificación.
5. Cuando se clasifican los datos de entrenamiento acorde a la construcción del árbol, se elige a qué nodo hoja será asignado cada uno de los datos, de modo que se respete la estructura del árbol.

Construcción de los árboles de decisión

Existen distintos criterios de selección de los atributos para la división óptima, todos tienen como objetivo encontrar los nodos más puros u homogéneos posibles en cada una de las ramificaciones. Los criterios se describen a continuación:

- **Error de clasificación**: Se define como la proporción de observaciones que no pertenecen a la clase más común en el nodo.

$$E_m = 1 - max_k(\hat{p}_{mk}) \tag{3.1}$$

- **Índice Gini**: Es una medida de la varianza total en el conjunto de las k clases del nodo m. Se considera una medida de pureza del nodo.

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.2}$$

- **Chi Cuadrado**: Esta aproximación consiste en identificar si existe una diferencia significativa entre los nodos hijos y el nodo parental, es decir, si hay evidencias de que la división consigue una mejora.

$$X^2 = \sum_{k} \frac{(observados_k - esperados_k)^2}{esperado_k} \tag{3.3}$$

- **Entropía**: Es una forma de cuantificar el desorden de un sistema. En el caso de los nodos, el desorden se relaciona con la impureza. Si un nodo es puro, es decir, contiene únicamente observaciones de una clase, su entropía es cero. Por el contrario, si la frecuencia de cada clase es la misma, el valor de la entropía alcanza el valor máximo 1.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk}) \tag{3.4}$$

3.5. Regresión Logística

Este algoritmo se usa comúnmente para estimar la probabilidad de que una instancia pertenezca a una clase en particular (por ejemplo, ¿cuál es la probabilidad de que este correo electrónico sea spam?). Si la probabilidad estimada es superior al 50 %, entonces el modelo predice que la instancia pertenece a esa clase, llamada clase positiva y etiquetada como "1", o bien, predice que no, es decir, pertenece a la clase negativa y etiquetada como "0". Esto lo convierte en un clasificador binario [17, Chapter 4].

Luego de haber utilizado el algoritmo previo para determinar las variables más importantes, se considera útil este algoritmo ya que permite estimar la probabilidad de que los repartidores vuelan a tomar una ruta, pudiendo así observarlas de forma gráfica, junto con poder clasificarlos según su propensión, siendo muy práctico para el objetivo del trabajo.

Algoritmo de la regresión logística

Un modelo de regresión logística calcula la suma ponderada de las características de entrada, más un término de sesgo, y genera la logística de este resultado.

$$\hat{p} = h_\theta(x) = \sigma(x^T\theta) \tag{3.5}$$

La función logística (σ) es una función sigmoidea (con forma de S) que genera un número entre 0 y 1, y definida como se detalla a continuación.

$$\sigma = \frac{1}{1 + exp(-t)} \tag{3.6}$$

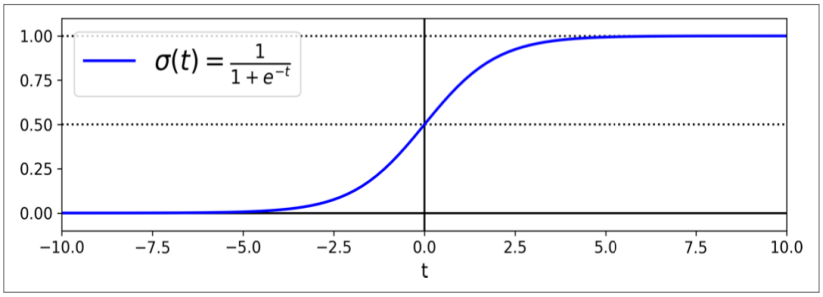

Figura 3.4: Función logística

Una vez que el modelo de regresión logística ha estimado la probabilidad $\hat{p} = h_\theta(x)$ de que una instancia "x" pertenezca a la clase positiva, puede hacer la predicción $\hat{y}$ fácilmente.

$$\hat{y} = \begin{cases} 0 & \text{si } \hat{p} < 0{,}5 \\ 1 & \text{si } \hat{p} \geq 0{,}5 \end{cases} \tag{3.7}$$

Dado que $\sigma(t) < 0{,}5$ cuando $t < 0$, y $\sigma(t) \geq 0{,}5$ cuando $t \geq 0$, un modelo de regresión logística predice 1 si $x^T\theta$ es positivo, y 0 si es negativo.

3.6. Métricas de desempeño de los algoritmos

A continuación se definen las métricas de desempeño, que suelen ser las más utilizadas, para determinar la capacidad predictiva del modelo que se desarrollará.

3.6.1. Matriz de confusión

Una matriz de confusión consiste en una representación matricial de los resultados de las predicciones de cualquier prueba binaria. Se utiliza a menudo para observar el rendimiento del modelo de clasificación de un conjunto de datos de prueba cuyos valores reales se reconocen. [17, Chapter 3].

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 3.5: Matriz de confusión

Cada predicción puede ser uno de los cuatro posibles resultados de la matriz, en función

de cómo coincide con el valor real. Para utilizar esta matriz se establece una de las clases como positiva y la otra negativa.

- **Verdadero Positivo (VP)**: Se predice como clase positiva y pertenece a esta.
- **Verdadero Negativo (VN)**: Se predice como clase negativa y pertenece a esta.
- **Falso Positivo (FP)**: Se predice como clase positiva y en la realidad pertenece a la clase negativa.
- **Falso Negativo (FN)**: Se predice como clase negativa y en la realidad pertenece a la clase positiva.

Teniendo en cuenta los conceptos previos, estos se relacionan con la prueba de hipótesis. Una hipótesis es una teoría basada en pruebas insuficientes que se intenta comprobar o rechazar mediante la experimentación u otros métodos. Luego de distintas pruebas una hipótesis puede ser considerada como verdadera o falsa. Una hipótesis que dice que no hay significancia estadística entre las dos variables de esta se conoce como Hipótesis Nula, y es la que el investigador pretende rechazar si tiene la suficiente evidencia para ello. Si bien las pruebas de hipótesis son confiables, hay dos tipos de errores que pueden ocurrir.

- **Error tipo I**: es equivalente a los Falsos Positivos (FP) y se traduce en rechazar una Hipótesis Nula que es verdadera.
- **Error tipo II**: es equivalente a los Falsos Negativos (FN) y corresponde a aceptar una hipótesis falsa nula.

A partir de la matriz de confusión, se construyen distintas métricas capaces de medir el poder predictivo del modelo:

- **Precisión**: indica el número de elementos clasificados correctamente en comparación con el número total de casos. Esta métrica no entrega información muy útil cuando los datos que se utilizan presenta clases desbalanceadas.

$$\frac{(VP + VN)}{(VP + VN + FP + FN)}$$

- **Sensibilidad**: indica la cantidad de verdaderos positivos que el modelo ha clasificado en función del total de valores de la clase positiva.

$$\frac{(VP)}{(VP + FN)}$$

- **Especificidad**: indica la cantidad de falsos negativos que el modelo ha clasificado en función del total de valores de la clase negativa.

$$\frac{(VN)}{(VN + FP)}$$

- **Valor predictivo positivo**: indica el número de la clase positiva predicha correctamente como una proporción del total de predicciones de la clase positiva realizadas.

$$\frac{(VP)}{(VP + FP)}$$

- **Valor predictivo negativo**: indica el número de clases negativas predichas correctamente como una proporción del total de predicciones de clases negativas realizadas.

$$\frac{(FP)}{(FN + FP)}$$

- **Prevalencia**: indica con qué frecuencia ocurre realmente la clase positiva en la muestra.

$$\frac{(VP + FN)}{(Total)}$$

3.6.2. Curva ROC y AUC

La curva ROC es un gráfico que representa el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Esta curva cuenta con dos parámentros.

- Tasa de verdaderos positivos (sensibilidad)
- Tasa de falsos positivos (1-especificidad)

Por otra parte, el AUC es el área bajo la curva ROC, y se interpreta como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio. Su valor oscila entre 0 y 1, siendo 1 un modelo cuyas predicciones son un 100 % correctas. Como se puede apreciar en la siguiente figura, cuando la curva ROC es la diagonal, el desempeño del modelo es el mismo que el que tuviese uno que clasificó de manera aleatoria. [17, Chapter 3].

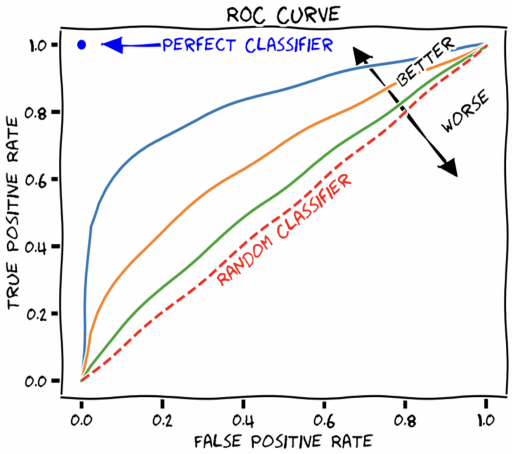

Figura 3.6: Curva ROC

Capítulo 4

Metodología

Para abordar el problema descrito previamente, se utilizará la metodología "Cross Industry Standard Process for Data Mining" (CRISP-DM). Se considera útil esta metodología dado que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, esta cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. El ciclo de vida del proyecto de minería de datos consiste en seis fases, las cuales presentan una secuencia flexible que permite movimiento hacia adelante y hacia atrás, y donde el resultado de cada fase determina qué etapa o tarea particular se debe realizar después.

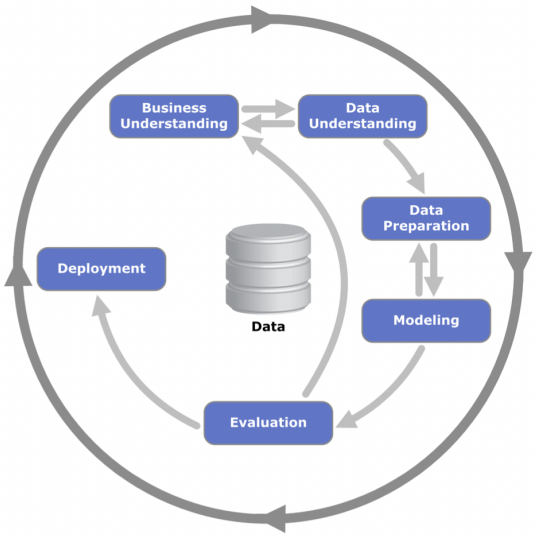

Figura 4.1: Diagrama metodología CRISP-DM

A continuación se detalla cada fase de la metodología que se utilizará:

1. **Comprensión del negocio**: En esta etapa inicial, el enfoque está en la comprensión de los objetivos del proyecto, definiendo las necesidades del cliente y la comprensión del negocio. Se busca convertir el conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar para alcanzar los objetivos definidos.
2. **Comprensión de los datos**: Esta fase comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar

los problemas de calidad, descubrir conocimiento preliminar sobre estos y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta. Los datos se obtienen de Google Cloud Platform (GCP), herramienta de Google utilizada por la organización para gestionar y almacenar su data.

3. **Preparación de los datos**: En la fase de preparación de datos se realizan todas las actividades necesarias para construir el conjunto final de datos, es decir, los que serán utilizados en las herramientas de modelado correspondientes al proyecto. Las tareas incluyen selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan. Esta etapa es desarrollada en Jupyter Notebook, aplicación web que facilita la creación e intercambio de documentos de programación y es usualmente utilizado para limpieza y conversión de datos.

4. **Modelamiento**: En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes a la oportunidad abordada, y se calibran sus parámetros a valores óptimos. Existen varias técnicas para el mismo tipo de problema de minería de datos, las cuales pueden tener requerimientos específicos sobre la forma de los datos, por lo que suele volverse a la fase de preparación de datos. La programación se realiza en lenguaje Python, y se emplea la librería Pandas, herramienta especializada en el manejo y análisis de estructuras de datos, junto con Scikit-learn, biblioteca de aprendizaje automático de software libre para el lenguaje de programación escogido. Posteriormente, se utiliza principalmente la curva ROC para determinar el desempeño de los modelos.

5. **Evaluación**: En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. La evaluación se realiza principalmente con métricas que se construyen a partir de la matriz de confusión. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

6. **Despliegue**: En esta última etapa, se busca organizar y presentar el conocimiento obtenido al cliente para poder utilizarlo, por lo que en este caso, lo que se realiza es una documentación detallada de las conclusiones obtenidas y cómo estas podrían utilizarse.

Capítulo 5

Desarrollo metodológico

5.1. Fase de comprensión del negocio

El objetivo de esta etapa consiste en comprender la organización, el área en que se va a trabajar, objetivos estratégicos, políticas, entre otros aspectos relevantes. A partir de esto, se espera definir el problema que se aborda en el proyecto y el objetivo de este. En este apartado se describe de forma detallada el problema, algunas definiciones relevantes para el trabajo y el proceso de generación de datos útiles para el desarrollo del proyecto.

5.1.1. Entendiendo el problema u oportunidad

Wareclouds es una startup que entrega servicios de fulfillment y última milla a los ecommerce, mediante un modelo colaborativo, donde son dos los actores principales que participan en la operación. Por una parte están los wareclouds, que son personas particulares que disponen de espacio en sus casas, donde almacenan y arman los pedidos de los ecommerce, mientras que los clouders, son repartidores que despachan estos pedidos a los consumidores.

La promesa de valor de esta empresa consiste en despachar los pedidos dentro de 24 horas en la Región Metropolitana. Para esto, Wareclouds en términos simples, todos los días genera distintas rutas, las cuales contienen todos los pedidos ingresados desde el día anterior después de las 11:00 hrs., hasta ese día antes de la misma. Luego de generar las rutas, estas son publicadas en la plataforma de los clouders, para que estos seleccionen la que más les interesa. Para que se pueda cumplir la entrega dentro de las 24 horas, todas las rutas diarias generadas deben ser asignadas a un repartidor.

A la escala actual, son cerca de 500 pedidos los que se deben despachar diariamente, donde cada ruta en promedio contiene 25 de estos, por lo que todos los días se tienen cerca de 20 rutas disponibles. El 80 % de estas rutas suelen ser tomadas de forma natural por los clouders, pero el resto debe gestionarse de forma manual por un trabajador de la empresa.

El proceso de venta manual de rutas, hoy consiste en que un trabajador debe comenzar a llamar a los repartidores que alguna vez han trabajado en Wareclouds, promocionándoles las rutas disponibles. La persona a cargo de esta actividad, intenta comunicarse primero con los clouders que han trabajado durante los últimos días y/o con los que observa que alguna vez han tomado una ruta similar, sin mayor información respecto a estos y sólo en base a su

juicio personal.

Wareclouds hoy tiene en promedio 100 pedidos mensuales retrasados porque las rutas a las que estos estaban asignados no lograron ser vendidas a tiempo, esto se traduce en que la persona que debe gestionar las rutas de forma manual no logra vender 4 rutas al mes. Considerando que el 20 % de las rutas diarias deben ser vendidas de forma manual (80 rutas mensuales), el 5 % de estas no logran ser asignadas.

Si bien la empresa hoy tiene en promedio 10.000 pedidos mensuales, debido a sus planes de crecimiento, espera duplicar su operación en julio y cuadruplicar los pedidos a fines de septiembre. Esto debido a que en mayo aterrizaría en México, donde existen cerca de 4 millones de Pymes, empresas que presentan problemas de transformación digital y logísticos, puntos en los que Wareclouds espera ser clave. Con esto también espera consolidar su operación para a fines de este año comenzar a operar en Estados Unidos. Junto con lo anterior, Wareclouds acaba de comenzar a ofrecer un nuevo servicio donde promete despachar en menos de 90 minutos, lo que también implicaría un aumento importante de clientes y pedidos mensuales.

En vista de lo anterior, con el crecimiento de los pedidos, también aumentarían las rutas diarias disponibles, lo que por consecuencia, genera más rutas que deban ser vendidas de forma manual. En el caso de julio, donde Wareclouds espera superar los 20.000 pedidos mensuales, se necesitarían vender cerca de 360 rutas de forma manual. Esto sería casi 5 veces las rutas que hoy se deben gestionar de esta forma. Es por esto, que implicaría tener a 4 personas más trabajando en este proceso. Bajo la misma lógica, en septiembre los pedidos serían cerca de 40.000, lo que se traduciría en 620 rutas que deban ser gestionadas, y que deberían tenerse entonces 8 trabajadores enfocados en esta labor.

Teniendo en cuenta que el 5 % de las rutas gestionadas por un trabajador mensualmente hoy no logran ser vendidas, esto implicaría que en julio las 5 personas, cada uno a cargo de 72 rutas mensuales aproximadamente, no lograrían asignar 18 rutas en promedio, lo que se traduciría en 450 pedidos retrasados ese mes. Por otra parte, en el caso de septiembre, donde serían más de 600 las rutas que no serían tomadas de forma natural por un clouder, bajo la misma línea, implicaría casi 800 pedidos retrasados debido a que estaban asignados a rutas que no lograron ser vendidas a tiempo.

Gestionar las rutas de forma manual y sin mayor conocimiento respecto al comportamiento macro de los clouders, dificulta entonces directamente la escalabilidad de la operación, donde primero, continuar realizando este proceso manualmente significaría agregar 4 trabajadores más para duplicar el número de pedidos, mientras que para cuadruplicarlos, se necesitaría en septiembre contar con 8 personas dedicadas a promocionar a los clouders las rutas que estén disponibles. Por otra parte, los resultados de este proceso hoy ya generan que existan pedidos que no cumplen con la promesa de valor de Wareclouds, que es entregarlos en menos de 24 horas. Mientras que acorde a las proyecciones de crecimiento, estos se cuadruplicarían con 20.000 pedidos mensuales y llegarían a ser cerca 800 en caso de cuadruplicar la operación.

En base a lo anterior, como solución a los problemas detallados previamente, se busca mejorar el proceso de asignación de las rutas que no son tomadas de forma natural por los clouders. Con foco en mejorar la asignación de rutas, primero se busca aumentar el conocimiento de los clouders que se inactivan y que posteriormente son notificados en busca de su reactivación, esto mediante un algoritmo de Machine Learning que permita conocer los atributos más importantes que determinan que los repartidores vuelvan o no a tomar una

ruta en Wareclouds. Teniendo esta información, luego se busca agrupar a los repartidores en función de su propensión utilizando un modelo de machine learning que finalmente permita proponer un piloto experimental que logre corrobar los resultados obtenidos y la propensión de reactivación, así como también permita aumentar la tasa de asignación automática de las rutas.

Finalmente, como resultado se espera que mediante el piloto experimental, al aumentar los clouders activos mediante la reactivación de estos, disminuyan las rutas que deban ser vendidas de forma manual, con lo que se requeriría menos personal para esta actividad, evitando ser un problema en la escalabilidad. También, disminuyendo las rutas que deban ser vendidas manualmente, disminuirían las rutas sin asignar, y con ello los pedidos retrasados, pudiendo así finalmente disminuir la fuga de marcas debido al incumplimiento en los plazos de entrega.

5.1.2. Los datos

En esta subsección se detalla cómo fueron generados los datos utilizados en el proyecto, los cuales consisten en enviar mensajes de WhatsApp de forma aleatoria a distintos clouders. El mensaje tipo se presenta a continuación.

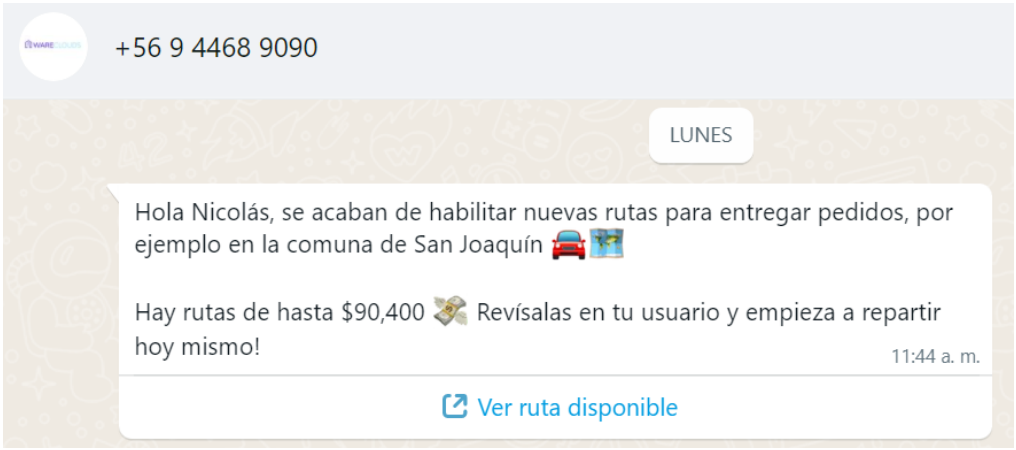

Figura 5.1: Mensaje de WhatsApp enviado a los clouders

Los mensajes contienen dos características principales. La primera, hace referencia a una de las comunas a las que se dirigen las rutas que posteriormente este clouder podrá observar en la plataforma. La segunda, es el precio que se muestra, el cual corresponde al precio más alto del conjunto de rutas que se encuentran disponibles.

5.2. Fase de recolección y comprensión de los datos

En esta etapa se recolectan los datos que serán utilizados posteriormente. El objetivo es comprender las bases de datos y evaluar la calidad de estas. Específicamente en esta sección se espera especificar las tablas con las que se trabajará y comprender las variables que existen en las bases de datos.

5.2.1. Base de datos

Los datos utilizados para llevar a cabo el desarrollo de este proyecto son los que contienen información de los mensajes enviados a los repartidores, junto con los que detallan la toma de rutas, esperando en estos ver el efecto de los mensajes en los clouders. También se consideran tablas que contienen características de los repartidores y de las comunas a las cuales se dirigen las rutas, en busca de generar una base de datos con variadas características. En total se utilizan siete tablas.

Las primeras dos tablas contienen información de los mensajes enviados. La primera corresponde a la de Mensajería para intentar que los clouders tomen alguna de las rutas promocionadas. Esta contiene 16.457 instancias y muestra los eventos de envío de mensajes, es decir, representa "Se envió un WhatsApp al número X a las HH:MM". La segunda tabla corresponde a la de Venta de rutas, la cual a diferencia de la anterior, presenta las características principales del mensaje enviado. Son 6.046 instancias que presenta esta tabla.

Una tercera tabla corresponde a las características de los clouders, la que contiene las principalmente variables relacionadas con información de contacto y dirección. Esta tabla presenta 468 instancias, las cuales corresponden a todos los clouders que han trabajado por lo menos una vez en Wareclouds.

La cuarta tabla contiene información de clouders que se les ha enviado algún tipo de mensaje, pero con un propósito distinto a reactivarlos, por lo que se considera importante para poder restar estos repartidores de las bases de datos de mensajería.

Otra tabla relevante es la que contiene todas las comunas de Chile, y no sólo los identificadores de estas, como se presentan en el resto de las tablas. Esta tabla es útil para obtener información de los sectores en los cuales viven los repartidores y en los que presentan mayor cantidad de repartos.

Finalmente, las últimas dos tablas presentan información respecto a las rutas. La primera contiene el historial de rutas tomadas los años 2021 y 2022, conteniendo cerca de 6.000 instancias. Mientras que la otra tabla relacionada, contiene las comunas a las cuales se dirigía cada ruta.

5.2.2. Atributos de la base

A continuación se presentan los datos recolectados previo a la limpieza y selección de los atributos. En los siguientes cuadros se pueden observar las tablas y variables de cada una de las fuentes de datos y el volumen de información que contiene cada una. A medida que se avanza en la metodología se detallará el proceso de limpieza y selección de variables, junto con explicar el significado de cada una de estas.

Tabla	**N° de instancias**
Clouders	468
Clouders activos	1.265
Comunas	348
Comunas de rutas	14.199
Historial de rutas	6.070
Mensajería	16.457
Venta de rutas	6.046

Tabla 5.1: Tablas utilizadas y su respectivo número de instancias

Tabla	**Variables**
Clouders	Id clouder
	Contraseña
	Nombre
	Apellido
	Número de identificación
	Dirección
	Detalles de dirección
	Id comuna
	Email
	Teléfono
	Licencia de conducir
	Patente
	Disponible
	Eliminado
	Fecha de creación
	Fecha de actualización
	Id tipo
	Tipo de clouder

Tabla 5.2: Tabla Clouders y sus atributos

Tabla	**Variables**
Clouders activos	Id clouder
	Tipo de clouder
	Fecha de creación
Comunas	Id city
	Nombre comuna
	Código zip
	Id región
Comunas de rutas	Id ruta
	Id comuna
Historial de rutas	Id ruta
	Id clouder
	Precio
	Fecha de creación
	Fecha de actualización
	Ruta optimizada
	Link Google Maps
	Link Waze
	Eliminada
	Distancia
Mensajería	Número remitente
	Número destinatario
	Id mensaje
	Estado del mensaje
	Cuerpo del mensaje
	Fecha
Venta de rutas	Id clouder
	Comuna promocionada
	Precio promocionado
	Rutas promocionadas
	Fecha

Tabla 5.3: Tabla Clouders activos, Comunas, Comunas de rutas, Historial de rutas, Mensajería y Venta de rutas con sus respectivos atributos

5.3. Fase de preparación de los datos

5.3.1. Limpieza de tablas

A continuación se detallan los criterios utilizados para limpiar las diferentes tablas previamente descritas, en función del objetivo del proyecto, se especifican los datos con los que se va a trabajar y se realiza el análisis exploratorio.

Comunas

- Dado que el trabajo se enfoca en las rutas y mensajes enviados en la Región Metropolitana, se eliminan las comunas que no pertenecen a esta región, quedando esta tabla con 53 filas.

Mensajería

- Considerando las variables que contiene esta tabla, el identificador del destinatario sólo es el número de teléfono de este, por lo que se comienza eliminando las instancias que no cuentan con esta variable, quedando 16.457 mensajes.
- Se eliminan los valores duplicados y quedan 16.455 mensajes.
- Cuando un clouder contesta el mensaje, el estado de este queda vacío, pero la variable "Cuerpo del mensaje" contiene lo que escribió, por lo que se eliminan los mensajes que tienen ambas variables vacías, asumiendo que este no fue recibido. También se remueven los mensajes con estado *failed*. Quedan 16.251 instancias.
- Luego de unir esta tabla con la de clouders, usando el número de teléfono como identificador, y eliminando los valores que no hacen match, es decir, no tienen un identificador común, quedan 16.045 filas.
- Teniendo el identificador del repartidor, se eliminan los que han sido notificados pero que no estaban inactivos, es decir, los que están en la tabla Clouders activos. Quedan 14.630 mensajes.
- Se eliminan también las instancias con email terminados en "@wareclouds.com", ya que son usuarios de la organización que han sido notificados para probar la función de mensajería. Luego de esto quedan 14.409 filas.
- Se agregan variables de la comuna del repartidor, uniendo la tabla con la base de datos Comunas. Posteriormente se eliminan las instancias que quedan con *id_city* vacío, dado que no pertenecen a la Región Metropolitana, quedando 14.351 instancias.
- Dado que esta tabla tiene más de una instancia por mensaje, es decir, registra cada estado por el que pasa el mensaje, se selecciona sólo el último estado para cada uno de estos. Quedan 5.320 filas.
- Finalmente, se remueven los mensajes que sólo estuvieron en estado *sent*, es decir, no llegaron al estado *delivered*. 5.105 filas quedan en la tabla Mensajería.

Venta de rutas

- Se eliminan los valores duplicado y quedan 6.031 instancias.
- Se une con la base de datos de Comunas, eliminando mensajes que promocionan rutas fuera de la Región Metropolitana, quedando 5.979 filas.

Teniendo ambas tablas de mensajería limpias (Mensajería y Venta de rutas), estas se unen creando la tabla "Eventos de mensajería", utilizando como variables en común la fecha de envío del mensaje y el destinatario, la cual finalmente contiene 4.987 filas.

Clouders

- Se eliminan las filas duplicadas, es decir, se eliminan las instancias que tienen el mismo número de teléfono pero distinto *id_clouder*. Quedan 461 repartidores.

Clouders activos

- Dado que esta base de datos contiene todos los mensajes enviados a los clouders activos, se eliminan los valores duplicados, siendo cada clouder una instancia, quedando 64 filas.

Comunas de rutas

- Se eliminan las rutas que están fuera de la Región Metropolitana, quedando 13.641 filas.
- Dado que en esta tabla cada instancia es una dupla ruta-comuna, la mayoría de las rutas se presenta más de una vez en la tabla ya que tiene asignada más de una comuna, por lo que se agrega la variable *sector*, para posteriormente determinar su sector predominante y dejar una ruta por fila. Quedan 5.751 instancias.

Historial de rutas

- Se eliminan rutas que están fuera de la Región Metropolitana, quedando 13.641 filas.
- Se eliminan valores duplicados e instancias que no tienen clouder asignado. Quedan 5.258 filas.
- Se une con tabla que contiene información de la comuna y sector de la ruta, y se eliminan valores que no tienen *id_city*, es decir, no están dentro de la Región Metropolitana. Quedan 4.994 instancias.

5.3.2. Consolidación de base de datos

En función de las tablas descritas anteriormente, se construyen dos tablas con la variable dependiente. Dado que el proyecto se puede abordar de dos formas distintas, se crean ambas tablas con distinto tipo de granularidad, es decir, una en función de los mensajes enviados y otra en función de los repartidos que se han intentado reactivar.

Efecto mensajes

La primera base de datos tiene enfoque en los mensajes, donde cada fila de esta es un mensaje distinto. La variable binaria de esta base, y futura variable dependiente, es si el mensaje tuvo o no un efecto en el clouder. Este atributo toma el valor 1 cuando posterior al mensaje, y en un rango de 15 días, dado que no se observan mayores reactivaciones posterior a ese período, el repartidor toma una ruta.

Esta base de datos nace de la unión entre Eventos de mensajería e Historial de rutas, teniendo como identificador común el id del clouder (*id_clouder*). Dado que existen valores que no hacen match, se eliminan esas instancias y esta tabla queda con 3.259 mensajes.

Efecto clouders

Utilizando la base de datos descrita previamente, se crea esta segunda tabla que en vez de estar en función de los mensajes, está en función de los clouders, es decir, cada fila es un clouder distinto. Tiene la misma variable dependiente que la tabla de mensajes. Esta base de datos cuenta con 140 clouders distintos. A continuación se detallan los atributos que contienen ambas tablas.

Tabla	Variables
Efecto mensajes/clouders	Id mensaje
	Estado del mensaje
	Cuerpo del mensaje
	Fecha mensaje
	Id clouder
	Fecha creación clouder
	Comuna clouder
	Sector clouder
	Comuna promocionada
	Precio promocionado
	Rutas promocionadas
	Sector promocionada
	Contador de mensaje
	Id ruta
	Precio ruta
	Fecha creación ruta
	Comuna ruta
	Sector ruta
	Diferencia mensaje-ruta
	Precio promedio clouder
	Días activos
	Número de rutas
	Frecuencia
	Sector de preferencia
	Sector clouder-ruta
	Diferencia de precio

Tabla 5.4: Tablas consolidadas y sus atributos

5.3.3. Creación de atributos

A continuación se detallan los atributos que fueron creados en las principales tablas para obtener variables explicativas para la fase de modelado.

Comunas

La primera creación de atributos que se realiza es en la tabla Comunas. El atributo "Sector" representa el sector de la Región Metropolitana al que pertenece la comuna, en busca de poder agrupar las rutas y los clouders en función del sector que tienen asociado, bajo la hipótesis de que este podría ser más relevante que las comunas en sí. Cada sector y las comunas que abarcan se detallan a continuación.

Sector	**Comuna**
Centro	Santiago
	Santiago centro
Chacabuco	Colina
	Lampa
	Tiltil
Cordillera	Pirque
	San Jose de Maipo
Maipo	Buin
	Calera de Tango
	Paine
Melipilla	Alhué
	Curacaví
	María Pinto
	Melipilla
	San Pedro
Nororiente	Las Condes
	La Reina
	Lo Barnechea
	Ñuñoa
	Providencia
	Vitacura
Norponiente	Cerro Navia
	Lo Prado
	Pudahuel
	Quinta Normal
	Renca
Norte	Conchalí
	Huechuraba
	Independencia
	Quilicura
	Recoleta

Tabla 5.5: Sectores y sus respectivas comunas

Sector	Comuna
Sur	El Bosque
	La Cisterna
	La Granja
	La Pintana
	Lo Espejo
	Pedro Aguirre Cerda
	San Bernardo
	San Joaquín
	San Miguel
	San Ramón
Suroriente	La Florida
	Macul
	Peñalolén
	Puente Alto
Surponiente	Cerrillos
	Estación Central
	Maipú
	Padre Hurtado
Talagante	El Monte
	Isla de Maipo
	Peñaflor
	Talagante

Tabla 5.6: Sectores y sus respectivas comunas

Clouders

En busca de obtener variables que permitan caracterizar a los clouders para poder determinar cuáles son más probables de reactivar, se crean distintos atributos en función de estos.

El primer atributo creado en esta tabla corresponde a una variable categórica llamada "Sector clouder" (*clouder_sector*), que representa el sector de la Región Metropolitana en el que vive el repartidor. Se observa que el conjunto de estudio de los repartidores pertenece principalmente al sector Nororiente de la capital (27 %), junto con el Centro (15 %) y el sector Sur (13 %).

Sector	Sector clouder
Centro	15 %
Chacabuco	1 %
Cordillera	0 %
Maipo	0 %
Melipilla	1 %
Norte	9 %
Nororiente	27 %
Norponiente	8 %
Sur	13 %
Suroriente	13 %
Surponiente	13 %
Talagante	0 %
Total	**100 %**

Tabla 5.7: Distribución de los sectores a los que pertenecen los clouders

La segunda variable creada en esta tabla corresponde a una variable categórica llamada "Sector de preferencia" (*route_clouder_preference*). Esta variable se construye a partir del historial de rutas tomadas por el repartidor cuando estaba activo, obteniendo así el sector de preferencia para trabajar. Se utiliza esta variable para el análisis exploratorio, específicamente para encontrar diferencias en el comportamiento de los repartidores en función de este atributo.

De las distribuciones por sector de preferencia se observa que los porcentajes varían de forma importante, siendo el sector Nororiente predominante (36 %), seguido por los sectores Sur (13 %), Norte (12 %) y Centro (11 %), respectivamente.

Sector	Sector de preferencia
Centro	11 %
Chacabuco	6 %
Cordillera	0 %
Maipo	0 %
Melipilla	1 %
Norte	12 %
Nororiente	36 %
Norponiente	6 %
Sur	13 %
Suroriente	9 %
Surponiente	6 %
Talagante	0 %
Total	**100 %**

Tabla 5.8: Distribución de los sectores de preferencia de los clouders

Luego se crea una tercera variable, que busca observar si influye el sector en el que vive el clouder sobre las rutas que toma, y por consecuencia, su sector de preferencia. La variable "Sector clouder-ruta" (*sector_clouder-route*) toma el valor 1 cuando estos dos sectores son iguales, y 0 cuando no.

De forma similar a la variable anterior se crean dos atributos categóricos que buscan representar si el sector del clouder y el de su preferencia se encuentran en la periferia de la Región Metropolitana. Las variables son "Provincia-sector clouder" (*province_clouder_sector*) y "Provincia-sector de preferencia" (*province_sector_preference*). Estas variables toman el valor 1 cuando el sector corresponde a Chacabuco, Cordillera, Maipo, Malipilla o Talagante.

Luego de tener caracterizados a los repartidores en función del sector en el que residen y en el que prefieren trabajar, se crean 4 variables continuas:

- Precio promedio clouder (*clouder_average_price*): promedio del precio de las rutas que el clouder tomó cuando estaba activo.
- Días activos (*active_days*): número de días que tomó rutas de forma frecuente antes de inactivarse.
- Número de rutas (*number_of_routes*): número de rutas que tomó antes de inactivarse.
- Frecuencia (*frequency*): número de días que estuvo activo dividido en el número de rutas que tomó durante ese período.

Eventos de mensajería

Análogo al atributo creado en la tabla Clouders, según la comuna promocionada del mensaje, se crea el atributo "Sector promocionado" (*prmoted_sector*), que corresponde al sector de la comuna que se le está promocionando en el mensaje. La distribución de los sectores se presenta a continuación.

Sector	**Sector promocionado**
Centro	7 %
Chacabuco	8 %
Cordillera	0 %
Maipo	2 %
Melipilla	2 %
Norte	23 %
Nororiente	12 %
Norponiente	14 %
Sur	17 %
Suroriente	9 %
Surponiente	3 %
Talagante	3 %
Total	**100 %**

Tabla 5.9: Distribución de los sectores promocionados a los clouders

Un segundo atributo creado en esta tabla es "Contador de mensaje" (*count*), que representa el número de mensajes que se le ha enviado a un determinado clouder, es decir, si esta variable toma el valor 10, significa que es el décimo mensaje enviado al repartidor.

Mensajes

El atributo que se crea en esta tabla, además de la variable dependiente, es la variable continua "diferencia de precio" (*dif_price*), la cual corresponde a la resta entre el precio promocionado en el mensaje y el precio promedio de las rutas tomadas por el clouder el tiempo que estuvo activo.

5.3.4. Variable dependiente

Luego de la creación de las variables mencionadas, se construye la variable dependiente "Efecto del mensaje" (*message_effect*). Dado que se tienen dos tablas de interés, se presentan los resultados obtenidos en función de los mensajes que hicieron efecto, así como los clouders que se activaron.

Mensaje	**Proporción**
Con efeto	3 %
Sin efecto	97 %
Total	**100 %**

Tabla 5.10: Distribución de la variable dependiente en la tabla Efecto mensajes

Clouder	**Proporción**
Con efeto	11 %
Sin efecto	89 %
Total	**100 %**

Tabla 5.11: Distribución de la variable dependiente en la tabla Efecto clouders

5.3.5. Análisis descriptivo de los datos

Se comienza realizando un análisis descriptivo en función de la variable "Sector", tanto el de preferencia de los clouders, como el de los promocionados. Si bien los repartidores muestran una importante preferencia por el sector Nororiente de la capital, los mensajes promocionados se concentran en este, pero también en la parte Norte y Sur de Santiago, y se observa que la mayoría de las reactivaciones se presentan en esos sectores. De aquí de forma previa se infiere que la cantidad de los mensajes promocionados para un determinado sector si influye en la reactivación, perdiendo en parte relevancia el sector en el que el repartidor prefiere trabajar.

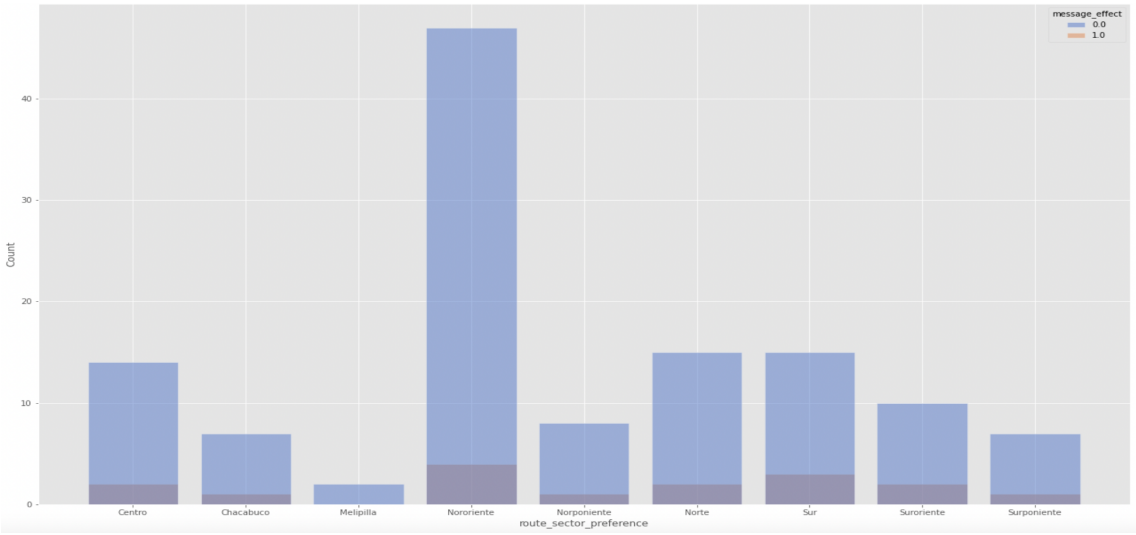

Figura 5.2: Distribución de los sectores de preferencia de los clouders

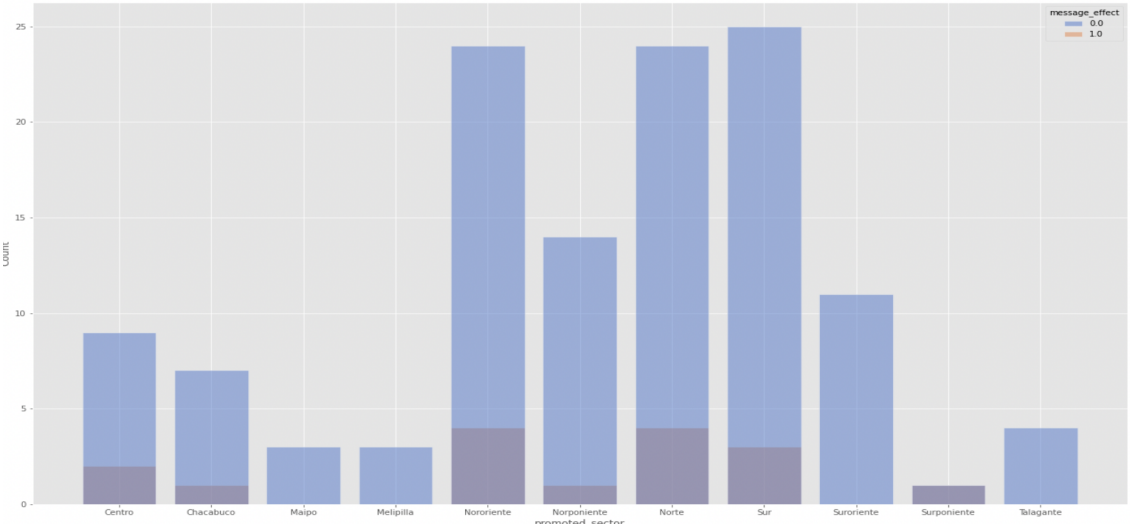

Figura 5.3: Distribución de los sectores promocionados en los mensajes

Luego se observan las principales variables continuas de los repartidores clasificadas según el sector de preferencia. En relación al valor promedio de las rutas que tomaron los repartidores estando activos, se observa que los clouders que prefieren el sector Norponiente presentan un menor precio promedio que el resto de los sectores de preferencia, así como también, que todos los repartidores que se reactivan han tomado rutas de menores valores en promedio que los que no vuelven a trabajar en Wareclouds. En concreto, el promedio del precio de los repartidores que no se vuelven a reactivar es de $36.848, mientras que el de los que vuelven a tomar rutas es de $28.824.

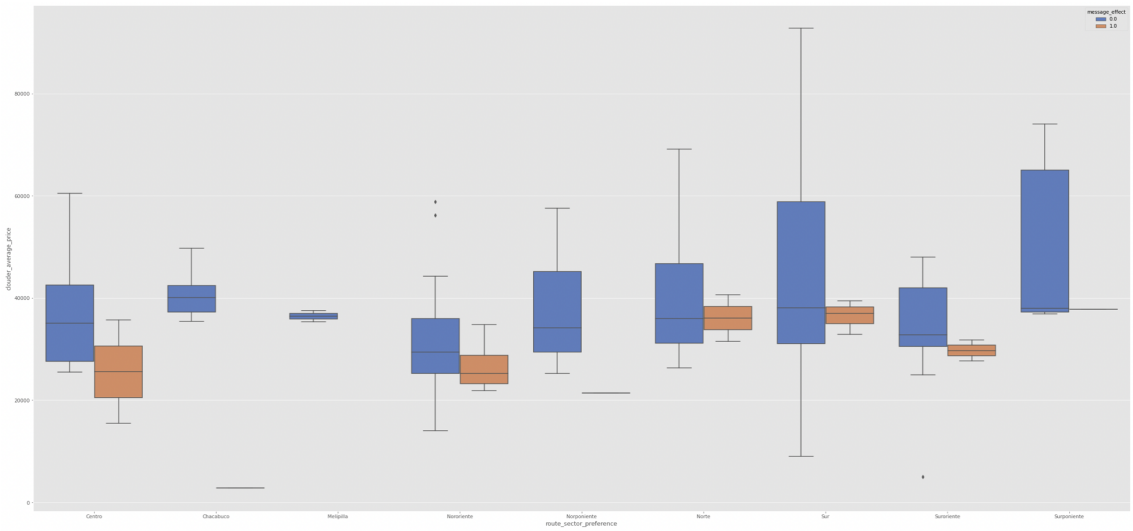

Figura 5.4: Precio promedio de las rutas tomadas cuando los clouders estaban activos categorizados por sector de preferencia

Por otra parte, respecto a los días que los clouders se mantuvieron activos, se observa una diferencia importante entre los que vuelven a trabajar en la organización y los que luego de 15 días no se asignan ninguna otra ruta. El promedio de días activos de los que se reactivan es más de 70 días, mientras que de los otros es sólo 35.

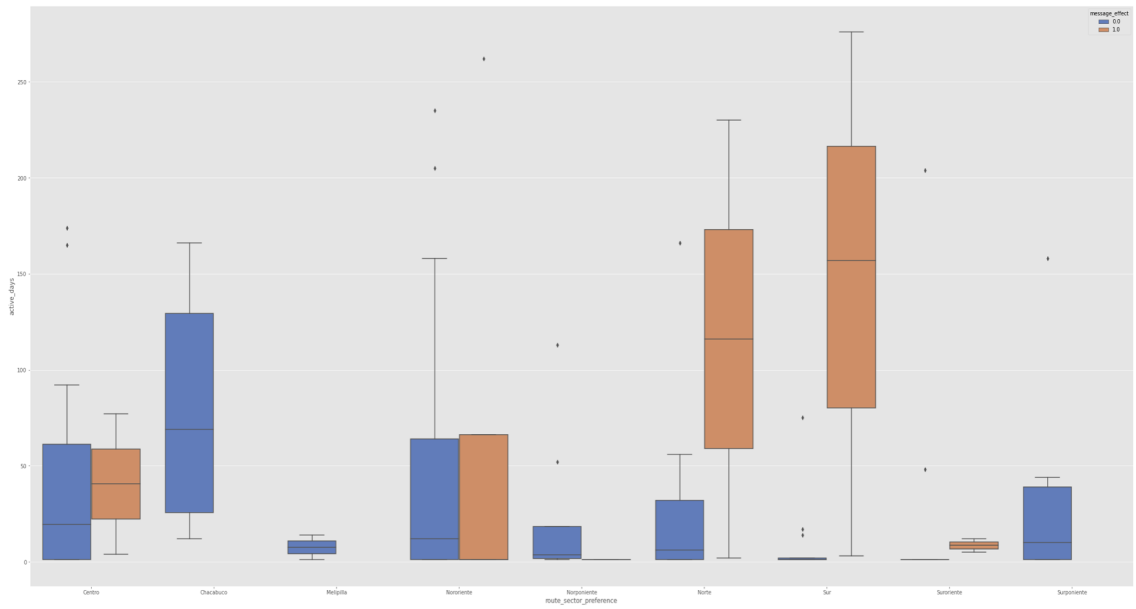

Figura 5.5: Días activos de los clouders categorizados por sector de preferencia

Respecto al número de rutas que los repartidores tomaron mientras estaban activos, el promedio de los que se reactivan es cercano a los 23, mientras que el de los que no, es sólo 8. En este atributo es importante destacar que este número se ve influenciado principalmente por los clouders que prefieren el sector Nororiente y Sur.

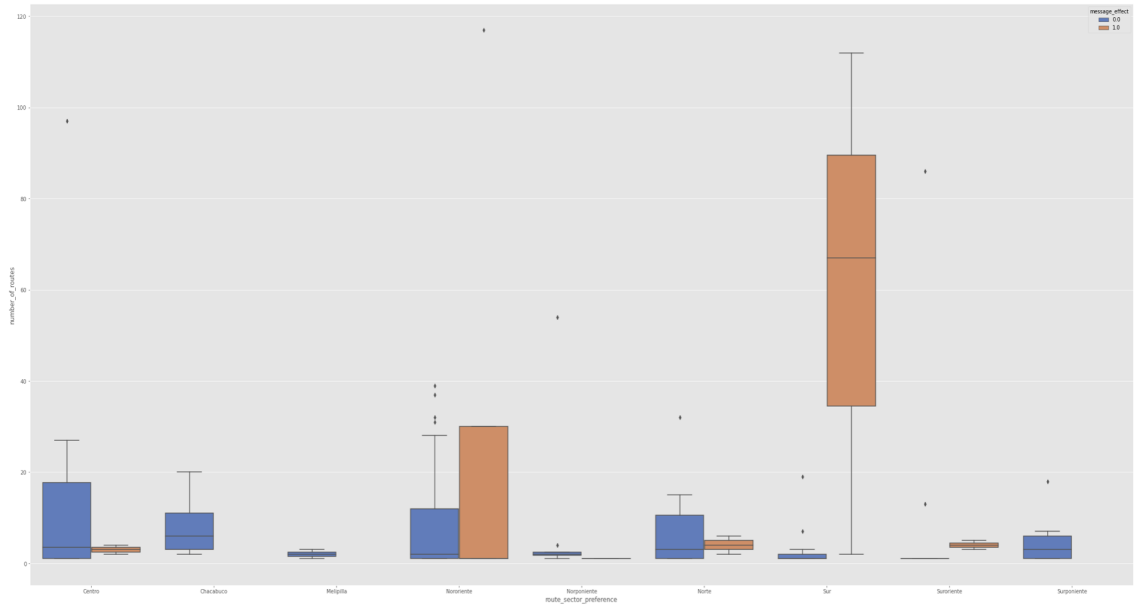

Figura 5.6: Número de rutas tomadas por los clouders categorizados por sector de preferencia

Una segunda parte del análisis exploratorio se realiza con las variables de los mensajes, y análogo al estudio hecho con los repartidores, se categorizan en función del sector promocionado. Una primera variable que se considera relevante es el precio, donde se tienen dos atributos importantes, el precio promocionado y la diferencia entre este y el precio promedio de las rutas tomadas por el clouder. Aquí se observa que en todos los sectores el precio

promocionado en los clouders que se activan fue menor que en los que no se activaron, sin embargo, es importante también notar que la variable "Diferencia de precio" es mayor. En concreto, la diferencia de los clouders que no se reactivan es en promedio de $11.990, mientras que en el caso de los que si lo hacen es de $13.119.

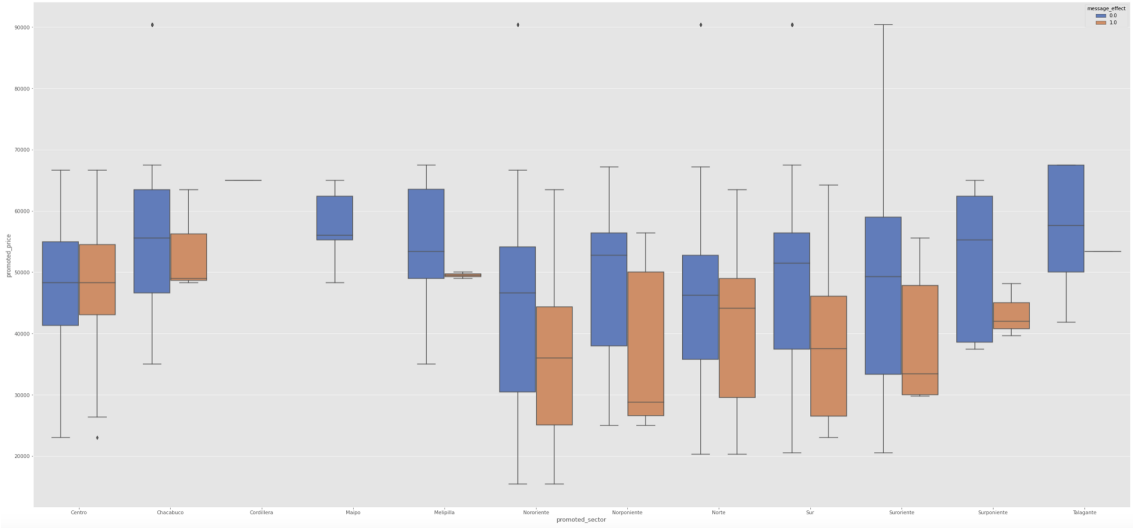

Figura 5.7: Precio promocionado de los mensajes categorizados por el sector promocionado

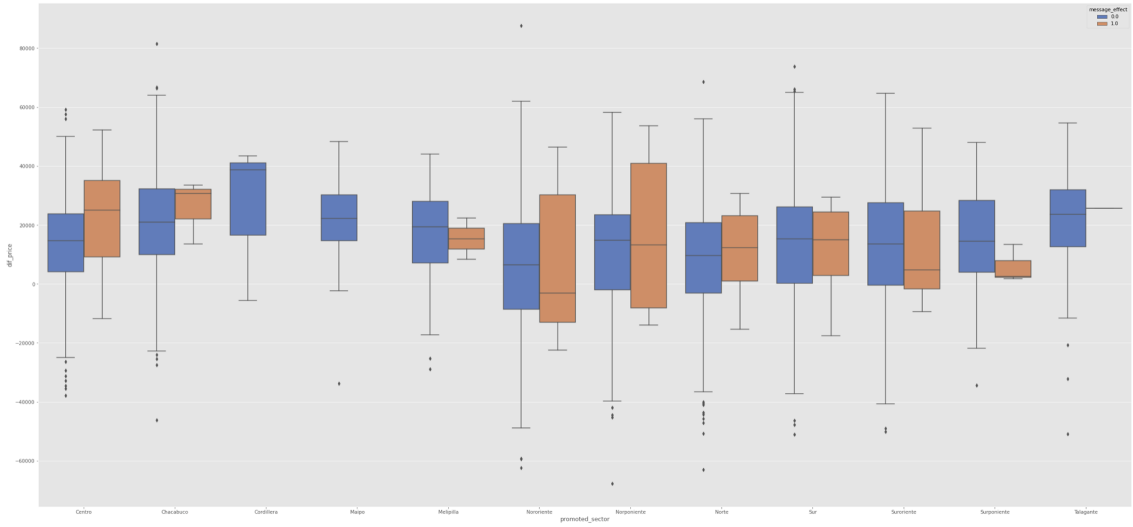

Figura 5.8: Diferencia entre el precio promocionado en el mensaje y el precio promedio del clouder categorizado por el sector promocionado

Finalmente, un último atributo que cobra relevancia en las características de los mensajes es el número del mensaje que se está enviando. Aquí se observa un importante efecto donde todos los clouders que se reactivan lo hicieron como máximo en el mensaje número 15, mientras que después de este no existe ninguna reactivación, y siendo en promedio el quinto mensaje el que logra que el clouder vuelva a tomar una ruta.

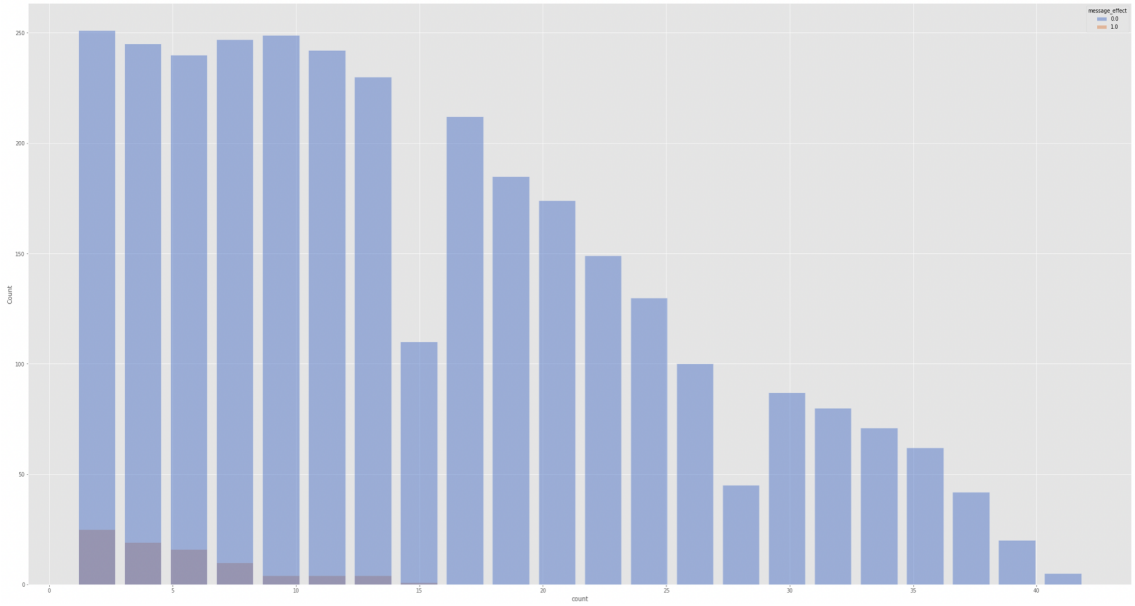

Figura 5.9: Distribución de la variable “Contador de mensaje”

5.4. Fase de modelado

Se utilizan dos algoritmos de Machine Learning para poder conocer los atributos más importantes que determinan la reactivación de los clouders, y posteriormente poder clasificarlos en función de su propensión para volver a tomar una ruta.

5.4.1. Machine Learning para determinar las variables más importantes en la reactivación de los clouders

Introducción

El uso de este algoritmo tiene por objetivo determinar los atributos más significativos de los clouders que influyen en su reactivación. Reactivarse en este contexto corresponde a volver a tomar una ruta dentro de los 15 días posteriores al envío de un mensaje.

Entrenamiento

Dada la naturaleza del problema la variable de interés "Efecto del mensaje" presenta un 11 % de casos donde los clouders se reactivan luego de un mensaje, por lo que existe un desbalance en la variable dependiente. Si bien el algoritmo no se utiliza con foco en la predicción, de igual forma implica que un predictor que clasifique a todos los clouders como caso positivo, es decir, que se reactive, tendrá en promedio un Accuracy de 89 %, lo que hace de esta medida una mala aproximación del desempeño. Por esto, para medir el desempeño, se proponen las métricas Area Under the Curve (AUC), Especificidad y Precisión, cuyas definiciones se encuentran en el capítulo 3.6 del Marco conceptual.

Se verifica la correlación entre las variables explicativas a partir de una matriz de correlación, con el objetivo de reducir las variables que estén altamente relacionadas y evitar considerar dos atributos como significativos y que finalmente sean redundantes. Se utiliza como punto de corte el valor 0,7 para considerar variables con alta correlación. En consecuencia, se observa que las variables más correlacionadas son: "Sector clouder" con "Comuna clouder", "Frecuencia" con "Número de rutas" y "Sector de preferencia" con "Provincia-sector de preferencia", respectivamente. Considerando lo anterior se dejan afuera los atributos "Sector clouder", "Número de rutas" y "Provincia-sector de preferencia". A continuación se presentan las variables que se consideran en el algoritmo.

Variables
Comuna clouder
Días activos
Frecuencia
Precio promedio clouder
Provincia-sector clouder
Sector clouder-ruta
Sector de preferencia

Tabla 5.12: Variables de los clouders que se consideran en el entrenamiento del modelo

Dada las posibles causas del problema y lo visto en el análisis exploratorio de los datos, la variable "Comuna clouder" se transforma en un valor numérico en función de la distancia a la que se encuentra determinada comuna del sector oriente, siendo mayor a medida que más lejos se encuentra de este. De forma similar, la variable "Sector de preferencia" toma el valor 0 cuando corresponde al sector oriente, y aumenta de forma discreta a medida que el sector de preferencia se encuentra a mayor distancia.

El modelo se entrena usando un 75 % de los datos y posteriormente se validan los resultados en el 25 % restante. Además, el entrenamiento se realiza ocupando la técnica de validación "K-fold Cross-Validation" con k=10. El resultado del algoritmo Decision Tree se presenta en la tabla 5.13.

Algoritmo	AUC	Accuracy	Sensitivity	Specificity	Precision (PPV)
Decision Tree	0,86	0,96	1,00	0,71	0,96

Tabla 5.13: Desempeño del algoritmo Decision Tree

Resultados

En la figura 5.10 se presenta la curva ROC del modelo, la cual tiene un AUC de 0,86, esto implica que con una probabilidad de 86 % el modelo podrá diferenciar a un clouder que se reactivará de uno que no.

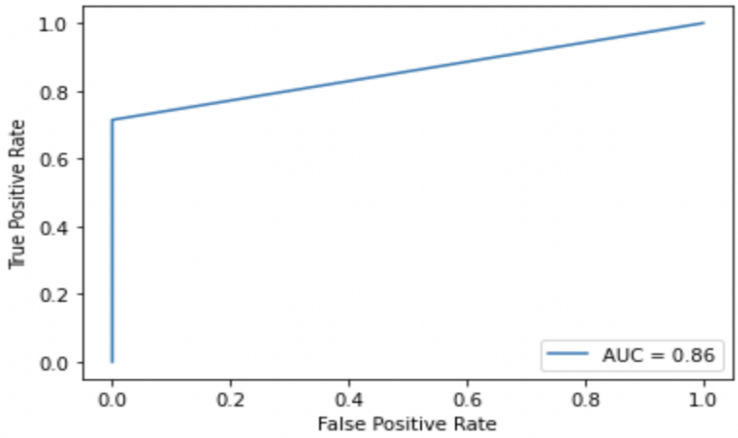

Figura 5.10: Curva ROC del algoritmo Decision Tree

El árbol de decisión indica las variables más relevantes a la hora de clasificar a los repartidores según su reactivación, las cuales son "Precio promedio clouder", "Días activos", "Comuna clouder", "Sector de preferencia" y "Frecuencia".

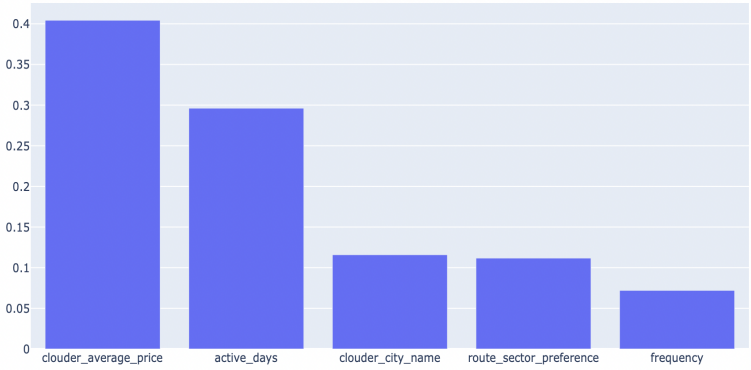

Figura 5.11: Importancia de variables del algoritmo Decision Tree

5.4.2. Machine Learning para conocer la propensión de reactivación de los clouders

Introducción

Se utiliza este algoritmo de regresión logística para determinar la probabilidad de reactivarse de cada clouder, y posteriormente poder clasificarlos según su propensión. Para este modelo se consideran las cinco variables más importantes determinadas previamente, las cuales se presentan en la tabla 5.14, con su respectiva importancia.

Variables	Importancia
Precio promedio clouder	0,404
Días activos	0,296
Comuna clouder	0,116
Sector de preferencia	0,112
Frecuencia	0,072

Tabla 5.14: Variables más significativas que determinan la reactivación de un clouder

Entrenamiento

El modelo de regresión se entrena usando también el 75 % de los datos y posteriormente se validan los resultados en el 25 % restante, pero a diferencia del modelo previo, sólo se utilizan los atributos más importantes. Para el problema de desbalance de datos se utiliza la técnica SMOTE, detallada en el capítulo 3.3 del Marco conceptual. Además el entrenamiento se realiza ocupando la técnica de validación "K-fold Cross-Validation" con k=10. El resultado del algoritmo se presenta en la tabla 5.13.

Algoritmo	AUC	Accuracy	Sensitivity	Specificity	Precision (PPV)
Logistic Regression	0,81	0,81	0,78	0,84	0,81

Tabla 5.15: Desempeño del algoritmo Logistic Regression

Resultados

En la figura 5.16 se presenta la curva ROC de la regresión, la cual tiene un AUC de 0,81, esto implica que con una probabilidad de 81 % el modelo podrá diferenciar a un clouder que se reactivará de uno que no.

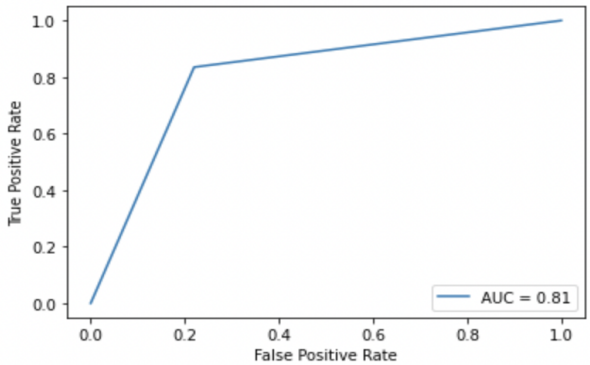

Figura 5.12: Curva ROC modelo Logistic Regression

Para observar como varían las propensiones de los 140 repartidores, se dividen estos en diez partes iguales, ordenados según la probabilidad de reactivarse, de manera que cada parte representa 1/10 de la muestra. El primer grupo presenta una probabilidad de reactivarse

cercana al 8 %, mientras que la del último decil es de 75 %. A continuación se observan de forma gráfica los 10 deciles y su respectiva propensión.

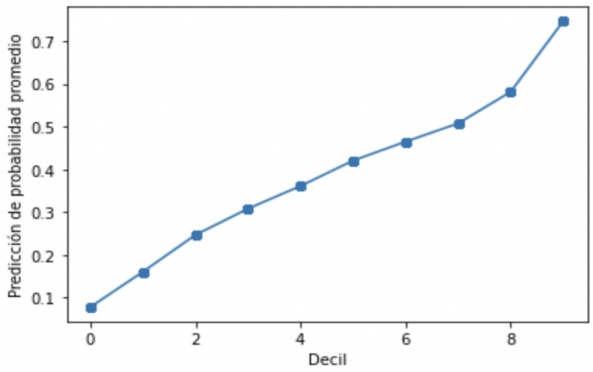

Figura 5.13: Curva de deciles y su respectiva propensión

De la mano con lo anterior, es también de interés saber cómo varían los valores de cada atributo según la propensión promedio de cada decil, por lo que se grafica la curva anterior con el valor cada atributo en el grupo respectivo.

Se puede interpretar la figura 5.14 de la siguiente manera:

- A medida que aumenta la propensión, disminuye en general el precio promedio de cada clouder, yendo desde $63.243 a $21.162. El único caso en el que a medida que aumenta la probabilidad no disminuye el precio es del decil 7 al 8, donde los valores son $26.205 y $26.271, respectivamente.
- Respecto a los días que estuvieron activos los clouders, a medida que aumenta la probabilidad promedio de reactivarse, aumenta en general el valor de este atributo. El primer decil tiene un valor promedio de 15,8, mientras que el último es de 110,6.
- Respecto al tercer y cuarto atributo, no se tiene una tendencia tan marcada, pero si se observa que en general los clouders con menor propensión viven y trabajaron en comunas más alejadas del sector oriente, mientras que los con mayor probabilidad de reactivarse viven más cerca de este sector y prefieren rutas cercanas a este.
- Respecto a la frecuencia, se observa que los clouders con menor propensión tomaron rutas de forma menos recurrente que los con mayor probabilidad de reactivarse. En concreto, la frecuencia del primer decil es de 7,7, mientras que del último es 2,3.

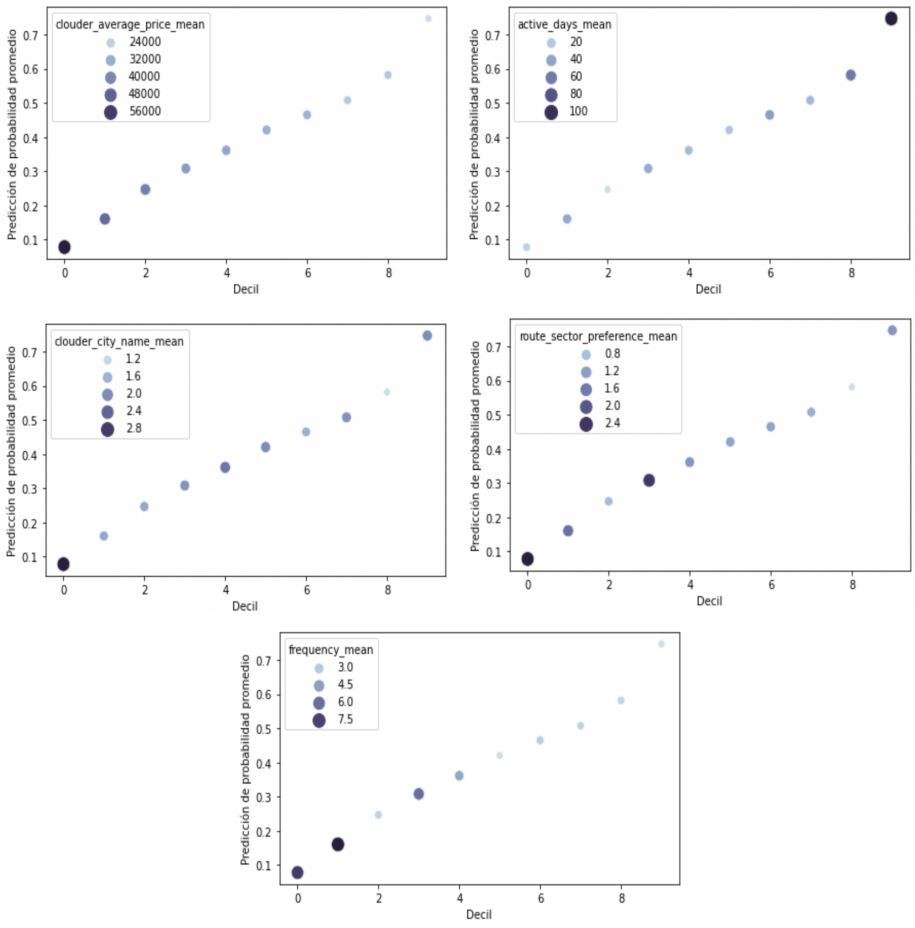

Figura 5.14: Curva de deciles y su propensión según cada variable

5.5. Fase de evaluación

Árbol de decisión

Para validar el desempeño del primer algoritmo se utiliza el 25 % de los datos que no se ocuparon durante el entrenamiento. El desempeño del modelo con estos nuevos datos es el siguiente:

- Accuracy: 91 %
- Sensitivity: 94 %
- Specificity: 50 %
- Precision: 91 %

Regresión logística

En el caso del modelo de regresión, se utiliza el 25 % de los datos que se dejaron aislados durante el entrenamiento para validar su desempeño, y posteriormente, se calculan los errores tipo 1 y 2, cuyas definiciones se encuentran en el capítulo 3.6 del Marco conceptual, y variando el peso de estos se determina el punto de corte en el que se considerarán los clouders con alta propensión.

El desempeño del conjunto de test es el siguiente:

- Accuracy: 83 %
- Sensitivity: 85 %
- Specificity: 50 %
- Precision: 83 %

Respecto a los errores tipo I y II, equivalentes a los Falsos Positivos y Negativos respectivamente, dado el tipo de problema que se está abordando, se considera que el error más costoso para la empresa es el de tipo I, donde se predice que un clouder se va a reactivar y realmente no lo hace. Dado el contexto del problema, sería más perjudicial este error ya que se tendrían considerados repartidores para una determinada fecha y estos finalmente no estarían disponibles para trabajar, afectando de forma negativa el cumplimiento de la promesa de valor de la empresa.

Teniendo en cuenta lo anterior, a continuación se observan los errores variando los puntos de corte de propensión, y a la vez aumentando el peso del error tipo 1 dentro del cálculo del error total.

	Error tipo I			
Punto de corte	50 %	60 %	70 %	80 %
0,5	8,6	9,7	10,9	12
0,55	7,1	8	8,9	9,7
0,6	5,7	6,3	6,9	7,4
0,65	4,3	4,6	4,9	5,1
0,7	4,3	4,6	4,9	5,1
0,75	2,9	2,9	2,9	2,9

Tabla 5.16: Valor del error según punto de corte de propensión y peso de Error tipo I

Dado que el error más bajo se encuentra en el mayor decil, se consideran los 14 clouders con mayor probabilidad los con alta propensión de reactivarse.

Para el caso de los otros repartidores, estos se clasifican según lo visto en las figuras 5.13 y 5.14. Dada la forma de la curva y los cambios en los valores promedio de los atributos más importantes, se consideran los primeros cuatro deciles como clouders con baja probabilidad de reactivarse, mientras que los que están entre el cuarto y décimo decil, excluyendo ambos, se clasifican como clouders con probabilidad media.

Teniendo ya definidos los tres grupos de repartidores según su propensión, se calcula el promedio de cada atributo en los respectivos grupos. En la tabla 5.17 se observan los resultados.

	Propensión		
Atributos	Baja	Media	Alta
Precio promedio clouder	46.883	29.654	21.162
Días activos	24	36	110
Comuna clouder	2,1	1,8	2
Sector de preferencia	1,9	1	1,3
Frecuencia	6,3	3	2,3

Tabla 5.17: Valor promedio de cada atributo según grupo de propensión

Se observa primero que a menor propensión, mayor es el precio promedio de las rutas que los clouders tomaron cuando estaban activos. De forma análoga, pero con una relación inversa, esto se presenta en el caso de los días activos y la frecuencia de los clouders que posteriormente se inactivaron, pudiendo identificar que los que tienen más probabilidad de reactivarse trabajaron en promedio más que los repartidores con menor propensión. Finalmente, respecto al sector de preferencia y comuna del clouder, no se observa una diferencia tan clara como en los otros atributos, pero respecto a la primera variable mencionada se puede ver que los con mayor propensión prefirieron sectores más cercanos al sector oriente, mientras que los con menor probabilidad despacharon en sectores más alejados. Finalmente, respecto a la comuna, se puede concluir que en general los clouders que se inactivaron vivían a distancias similares del sector oriente.

5.6. Fase de despliegue

5.6.1. Productivización del modelo

Finalmente con los conocimientos obtenidos de las fases previas, se propone un piloto que permita corroborar el comportamiento de los clouders inactivos, aumente el número de reactivaciones y sea una base para próximos proyectos relacionados con la problemática presentada. El experimento propuesto se enfoca en variar determinadas condiciones de los mensajes y luego ver cómo se comportan los tres grupos de repartidores, permitiendo afirmar o refutar las hipótesis de su conducta.

Dada las características más relevantes que se determinan, y considerando que el precio promedio de las rutas que tomaron los clouders cuando estaban activos es el atributo más significativo en la reactivación, la primera variable que se considera importante añadir está relacionada con un incentivo económico hacia estos. Es así como se propone que en el mensaje se ofrezca un bono adicional por adjudicarse alguna de las rutas promocionadas.

Ya que a medida que aumenta la probabilidad de reactivarse, disminuye el valor promedio de la variable "Precio promedio clouder", se espera que agregando el incentivo económico aumente el porcentaje de reactivación de los clouders con menos propensión, dado que por lo visto estos serían más susceptibles a una retribución económica y podrían presentar mayor interés mientras mayor sea el pago que reciban.

Una segunda variable que se considera de interés modificar es la frecuencia en la que se envían los mensajes, ya que dada las diferencias de la periodicidad en la toma de rutas de los distintos grupos, se cree que esto podría deberse a la menor frecuencia en la que se presentan rutas de interés para los clouders con menor propensión, y que finalmente provoca su inactivación. Es así como esto podría ser un tipo promoción directa de rutas, facilitando el ingreso de estos a la plataforma y pudiendo observar si existen rutas de su interés, tanto en términos de destino y/o de valores de estas.

Modificando la frecuencia con la que se ofrecen nuevas rutas a los repartidores, se esperaría observar un aumento en el número de reactivaciones, tanto de clouders con propensión media y baja. En este atributo, existe el riesgo de que los repartidores con menor propensión presenten una menor periodicidad en el trabajo porque realmente les acomoda más esta modalidad, y podrían considerar molestas las notificaciones, pero se espera de todas formas que en los de propensión media si exista un efecto positivo.

Finalmente una última variable que se propone modificar es la comuna promocionada, y en este caso cambiarla para que en el mensaje se especifique el sector de las rutas que se le están ofreciendo al repartidor. Esto ya que el atributo "Sector de preferencia" es uno de los más relevantes, y dado que existe una diferencia entre los clouders de baja propensión con los de media y alta.

Llevando a cabo esta modificación, se esperaría primero observar un aumento de reactivaciones, principalmente del grupo con menor probabilidad de volver a trabajar en Wareclouds, cuando el sector esté más alejado del sector oriente. También, dado que un sector contiene distintas comunas, se esperaría que implementando esto aumenten por lo menos las aperturas en la plataforma para observar las rutas disponibles, pudiendo aumentar así también las reactivaciones en los tres grupos, ya que no se quedarían sólo con la idea de la comuna promocionada que actualmente se observa en el mensaje. Finalmente, la idea de este cambio

también sería observar de forma más detallada cuales son las comunas de mayor interés en cada grupo, para obtener más información de estos, y en un futuro poder continuar clasificándolos, buscando que la mensajería sea aún más focalizada.

5.6.2. Impacto económico

Para estudiar el impacto económico que tendrían los resultados del trabajo y la posterior implementación del experimento propuesto, se estima el número de reactivaciones que existirían en tres escenarios: optimista, medio y pesimista. Teniendo en cuenta que hoy son 320 las rutas que se asignan de forma automática, en promedio cada clouder toma 8 rutas al mes. En función de esto, se estima el impacto que tendría cada clouder que se vuelve a activar en las rutas que deberían ser vendidas de forma manual en el período de tiempo proyectado.

Escenario optimista

El mejor escenario considera reactivar a 84 repartidores, que corresponden a los con media y alta propensión, y que se traduciría en un aumento promedio de 672 rutas mensuales asignadas de forma automática.

En términos de requerimientos de personal, considerando los nuevos clouders activos, y así la disminución de rutas que deban ser vendidas de forma manual, en este caso no se requeriría contratar a ningún trabajador adicional, ya que en el período estudiado, el número máximo de rutas que deban ser promocionadas de forma manual es 620. Teniendo presente que el costo empresa de la persona que realiza esta labor es $820.000, y asumiendo que esta persona se mantendrá en la organización, el ahorro que se tendría en este escenario es un poco más de $16.000.000, lo que corresponde en términos porcentuales a 77 %.

El segundo impacto económico está relacionado con la fuga de clientes debido a los pedidos retrasados por rutas no tomadas a tiempo. La salida de las marcas por disconformidad en tiempos de entrega es hoy cercana al 2 %, considerando que son en promedio 3 marcas las que se fugan mensualmente por este problema. En base a las proyecciones de crecimiento, sin implementar ninguna mejora en la venta de rutas, el costo de la salida de ecommerce se traduciría un monto cercano a los $184.000.000

Considerando lo planteado previamente, con la implementación del experimento propuesto en el mejor escenario, dado que se asignarían todas la rutas de forma automática, se tendría un ahorro del 100 % del costo mencionado.

Un último impacto económico está relacionado con eventos que aumentan de forma importante las ventas de los ecommerce, como los son los Cyber Days o festividades como Navidad. Al comienzo del informe se comenta que en un día normal suelen asignarse de forma automática el 80 % de las rutas, mientras que cuando aumentan los pedidos esta suele bajar en promedio en un 20 %. También se describe que han existido instancias donde debido a falta de repartidores en estas fechas, marcas importantes han decidido desligarse de la empresa por no poder cumplir con la promesa de valor, lo que se ha traducido en un costo promedio de $3.300.000 por evento.

Para el cálculo de este tercer impacto económico se estiman los pedidos que habrían en un año, teniendo presente el período estudiado, y se determina que en promedio serían 1.120 diarios. Por otra parte, se determina que en promedio en los días de *peaks* aumentan por lo

menos un 60 % los pedidos. Esto finalmente se traduce en que existirían 72 rutas diarias y 29 de estas deberían ser vendidas de forma manual sin implementar ninguna mejora en el proceso.

Teniendo en cuenta lo descrito previamente, y asumiendo que en este escenario son 84 los repartidores adicionales que estarían activos, y dado que en estas fechas la mayoría de los repartidores, debido a los incentivos que existen y al aumento del valor de las rutas, se encuentran disponibles para trabajar, no quedarían rutas pendientes por tomar, por lo que el ahorro del costo que genera la fuga de marcas en este caso es del 100 %.

Escenario medio

En este escenario, se espera reactivar 49 clouders, que corresponden a la mitad de los con propensión media, junto con la totalidad de los con mayor probabilidad. Esto se traduciría en un aumento promedio de 392 rutas mensuales asignadas de forma automática.

En relación a la contratación de nuevas personas para realizar la venta manual de rutas, en este caso si se requeriría de más trabajadores en los últimos meses del período considerado, y lo que en este escenario implicaría una disminución cercana a $14.000.000.

Respecto a la fuga de clientes debido al incumplimiento en los tiempos de entrega, en este escenario, la disminución de pedidos retrasados implicaría un ahorro cercano a $140.000.000, que equivale a un 76 % del costo que se tendría sin realizar ninguna mejora en el proceso.

Como último impacto, análogo a lo detallado en el mejor de los casos, en este escenario son 49 los repartidores que se estiman que estarían disponibles para trabajar, por lo que tampoco se incurriría en el costo de los cerca de $17.000.000 que provoca la fuga de estas marcas en un año.

Escenario pesimista

Finalmente, en el peor escenario, sólo se considera la reactivación de los con alta propensión, es decir, 14 clouders. Esto entonces implicaría un aumento promedio de 112 rutas mensuales asignadas de forma automática.

De forma análoga a los escenarios anteriores, comenzando con el impacto económico que tendría la disminución de nuevo personal para esta labor, esto se traduciría en una disminución de un poco más de $4.000.000, y así un ahorro en términos porcentuales de casi un 20 %.

Siguiendo con la fuga de clientes, en el escenario con menor impacto económico, se estima que el efecto del proyecto en la salida de marcas por incumplimiento de tiempos de entrega significaría un ahorro de $40.000.000, lo que corresponde una disminución de más del 20 %.

Finalmente, respecto al último impacto económico estudiado, en este escenario, debido a que serían 14 los clouders que se estiman volverían a trabajar en Wareclouds, a diferencia de los escenario previos, en este caso, se estima que quedarían 15 rutas por vender manualmente. En función de esto se determina que serían 2 las marcas que se fugarían en cada evento, por lo que el ahorro sería de un 33 %, que se traduce en cerca de $6.000.000.

5.6.3. Riesgos del proyecto

Los riesgos del proyecto están relacionados principalmente con la última etapa de este, donde se propone implementar un piloto experimental que permita corroborar los resultados obtenidos y las inferencias que se realizan de estos. El primero se relaciona con el tiempo y los recursos de la empresa que serían utilizados en este. Si bien el período de implementación no debería ser extenso, de igual forma se estarían focalizando los esfuerzos en esto y no en otro proyecto, el cual podría tener mayor urgencia y/o resultados. Tanto el tiempo como los recursos de Wareclouds cobran mucha relevancia ya que es una startup, como se contextualiza al inicio del trabajo, por lo que el número de trabajadores es reducido, y dada la etapa en la que se encuentra la organización, existen variadas oportunidades de mejora en distintos ámbitos y con niveles de prioridad similares.

Considerando lo anterior, en este caso existe un *trade-off* donde se podrían presentar beneficios económicos, junto con mejorar la experiencia de los clientes y de los consumidores con el despacho, pero existiendo falencias en otros procesos de la organización a los que se estaría renunciando. Para este caso en particular se propone un período de prueba acotado, donde se observen los resultados del piloto, y en caso de ser positivos continuar con el proyecto. En caso contrario, focalizar los recursos de la empresa en otros proyectos.

Por otro lado, el proyecto presenta el riesgo de no obtener los resultados esperados, lo cual podría deberse a la cantidad de datos que se utilizaron en la fase de modelado. También debido a que es la primera oportunidad en la que se estudia el comportamiento de un grupo de clouders mediante datos, por lo que no existen conocimientos previos en función de estos que respalden o guíen las hipótesis planteadas. En este caso, podría no presentarse el aumento esperado de reactivaciones con los incentivos propuestos y por lo tanto, continuar existiendo retrasos en las entregas y los efectos negativos que eso conlleva. En este caso, lo ideal sería pausar el piloto esperando recolectar una mayor cantidad de datos. Luego de eso, realizar la limpieza de estos y correr nuevamente los algoritmos con foco en sus resultados.

Capítulo 6

Conclusiones

6.1. Cumplimiento de objetivos

De acuerdo a los objetivos planteados en la investigación, y una vez analizados los resultados, se presentan las siguientes conclusiones.

Con relación al primer objetivo del estudio que es Identificar las reactivaciones que existen posterior al mensaje, el cual se aborda de dos formas, una en función de los mensajes que generan reactivaciones, y otra en función de los clouders que se reactivan, se obtiene que en ambos casos estas son escasas, donde en el primero no superan el 3 %, y en el segundo son el 11 % de los clouders. Sin embargo, este fenómeno se explicaría dada la aleatoriedad en la que están programados los mensajes, así como también debido al período acotado en el que se estudiaron las reactivaciones, dado el plazo que el proyecto presentaba.

Para el segundo objetivo relacionado con Determinar las variables más importantes que influyen en la reactivación de un clouder, se observa que si bien son cinco las variables más importantes, son dos las con mayor relevancia y que presentan una clara relación entre la probabilidad de reactivación y el valor de estas variables. El atributo que lidera en términos de importancia es el precio promedio de las rutas que se asignaron los repartidores cuando estuvieron activos, y muestra que a medida que aumenta la probabilidad, disminuye este valor. De esto se concluye que existiría una diferencia importante en el interés de los repartidores que se buscan reactivar, donde los menos propensos a esto presentan una mayor inclinación a la retribución económica y buscan que el pago sea considerablemente más alto que el resto.

La segunda variable de mayor relevancia corresponde a los días que trabajaron los clouders mientras estuvieron activos, presentando una relación directa entre estos y la probabilidad de reactivación. En función de lo anterior, se concluye que el grupo con mayor propensión estaría interesado en trabajar de forma más constante, mientras que los con menor propensión es probable que vean esta oportunidad como algo más esporádico y sean útiles para la organización en momentos de *peaks* de pedidos, más que buscar que estos trabajen de forma constante.

Para el objetivo Clasificar a los clouders según su propensión de volver a tomar una ruta, se observa que sólo el 10 % puede ser clasificado como de alta propensión y que este grupo presenta valores muy característicos en los atributos más relevantes. De aquí se destaca que podría existir la posibilidad de anticiparse a que estos repartidores se inactiven, personalizan-

do de forma previa beneficios para estos. Por otra parte, también se observa que la mayoría de los clouders presentan propensión media, y dado que contiene a la mitad del grupo de estudio, tiene una alta desviación en los valores. Considerando lo anterior, es esperable que frente a los incentivos, un número considerable de ellos responda de forma positiva a estos.

Finalmente, para el último objetivo que corresponde a Proponer un diseño experimental para validar la hipótesis de mensajes y propensión, se plantean tres variables que permitirían corroborar los resultados obtenidos en el trabajo, así como también mediante su implementación disminuirían el costo económico en caso de que el proceso no experimentara ninguna modificación, obteniendo resultados que aumenten la escalabilidad de este y contribuyan con el crecimiento de la empresa.

Las modificaciones en las variables propuestas están enfocadas principalmente en los grupos de menor propensión según sus características más importantes, sin embargo, no serían negativas para los clouders con mayor probabilidad de reactivarse. En función de lo anterior, además de robustecer los resultados obtenidos, presentándose las reactivaciones esperadas, se obtendría un impacto importante en el cumplimiento de la promesa de valor.

Teniendo en cuenta lo anterior, son principalmente tres acciones que se proponen en el piloto experimental para un trabajo futuro en base al conocimiento obtenido. Primero, considerando la variación que existe en el precio promedio de las rutas tomadas en los tres grupos de repartidores, incluyendo un incentivo económico en la mensajería, aumentaría de forma significativa la reactivación. Segundo, dada la diferencia de periodicidad en el trabajo, se propone modificar la frecuencia en la promoción de rutas, aumentando así las vistas de estas. Y como última variable, se debería especificar en el mensaje el sector de las rutas promocionadas, ya que el sector de preferencia es una variable determinante en los grupos de clouders.

6.2. Hipótesis planteadas

Respecto a la primera hipótesis presentada al comienzo del trabajo, relacionada con la desproporción en la frecuencia de los destinos, lo que generaría desinterés en los repartidores que prefieren sectores más alejados del sector oriente, se observa que esta si sería una causa para que se inactiven. Si bien no es muy claro en los clouders con mayor probabilidad de volver a trabajar en la organización, si se visualiza que el grupo con menor propensión es el que presenta mayor interés en sectores más alejados de la zona nororiente, por lo que ya que estos se publican de forma menos constante, dejarían de interesarse en el trabajo en Wareclouds.

En relación a la segunda hipótesis planteada, enfocada en el tipo de comunicación con el que se notifican las rutas diarias disponibles, no se obtienen resultados vinculados a esta, por lo que no se considera adecuado afirmarla ni refutarla. Lo que si, dado los conocimientos obtenidos del proyecto, sería interesante estudiar esta variable posteriormente, ya que dada la diferencia de interés en la organización y la forma de trabajar de los clouders, esta podría estar relacionada con el tiempo que toman en observar las rutas.

Finalmente, respecto a la última hipótesis relacionada con el tipo de interés en el trabajo y los incentivos de los clouders, se observa que si existirían diferencias entre los grupos de repartidores estudiados respecto a la modalidad de trabajar, pero esto podría deberse a la

menor frecuencia en la que se presentan rutas más alejadas del sector oriente, que son las que prefieren los repartidores con menor probabilidad de reactivarse, e implicaría que los clouders de menor propensión trabajen de forma más esporádica. Respecto a los incentivos, si existe una diferencia sustancial entre los valores de las rutas tomadas por cada grupo, por lo que en este caso incentivos económicos si se esperaría que tuvieran un efecto importante, sobre todo en los repartidores de menor propensión.

6.3. Impacto del trabajo realizado

Respecto al impacto general del trabajo realizado, el primero está relacionado con la dificultad de escalabilidad que presenta hoy el proceso estudiado. Como es descrito al comienzo, si la venta de rutas manual se mantiene tal como está, esto implicaría contratar a cuatro personas más para duplicar la operación, y agregar nuevamente a tres trabajadores más para cuadruplicarla.

Teniendo en cuenta el conocimiento adquirido del trabajo y llevando a cabo el piloto experimental propuesto, se podrían disminuir el número de personas necesarias en más de la mitad en el período estudiado, necesitando así sólo contratar a tres trabajadores más en los 6 meses. Teniendo lo anterior en consideración, el trabajo contribuye de forma importante en disminuir la dificultad del proceso sobre el escalamiento de la operación, requiriendo menos personal a medida que la empresa crece.

Un segundo impacto está relacionado con la fuga de marcas debido al incumplimiento en los plazos de entrega. Este es un tema muy relevante, dado que si bien tiene un costo económico importante, también tiene un gran costo en la reputación y crecimiento de Wareclouds.

Dado que la razón de la salida de los ecommerce es debido al incumplimiento de la promesa de valor, las marcas se van con una percepción muy negativa de la empresa, que afecta directamente una de las métricas importantes de la organización conocida como NPS (Net Promoter Score), mencionada en la sección de Desempeño Organizacional. Este es un indicador de la experiencia del cliente y mide las probabilidades de que estos recomienden la empresa.

En relación a lo anterior, el impacto de disminuir la fuga de marcas por esta razón, permitiría aumentar la reputación y experiencia en Wareclouds, junto con la métrica mencionada, contribuyendo así a la factibilidad de los planes de crecimiento de la empresa.

Teniendo en cuenta los principales impactos mencionados previamente, junto con el impacto económico de estos, en promedio, el trabajo en su conjunto permitiría disminuir casi en un 60 % las personas necesarias para trabajar durante el período de tiempo estudiado, así como también permitiría disminuir en un 65 % la fuga de marcas por incumplimiento en la promesa de valor.

Finalmente, el impacto económico total estimado puede ir entre $216.000.000 y $49.000.000, permitiendo un ahorro promedio de $145.000.000. Esto finalmente se traduce en una disminución del 66 % de los costos que se tendrían al no realizar ninguna acción de mejora en el proceso estudiado.

6.4. Cambios en la metodología

Durante el desarrollo del proyecto, debido a los datos observados y el resultado de estos, se decidió modificar lo que se realizaría en la fase de modelado y de despliegue del trabajo. En un comienzo el objetivo de este proyecto sería establecer un modelo de predicción que permitiera determinar, en base a los mensajes enviados y al clouder, si este se reactivaría o no. Debido a que se observaron pocas reactivaciones en el período de tiempo considerado, y teniendo en cuenta que no había ningún estudio previo acerca de los repartidores, se optó por utilizar Machine Learning, pero con el objetivo de aumentar el conocimiento de los repartidores y poder clasificarlos según su propensión, para proponer finalmente un piloto experimental que mejore el proceso de venta de rutas.

Debido a la cantidad reducida de datos disponibles que se fijaron en base a los plazos del proyecto, y dada la complejidad de un modelo de predicción, se consideró que diseñar un modelo enfocado en predecir qué clouders se reactivarían en función de sus características y de los mensajes recibidos, podría no presentar un buen desempeño, y a la vez tampoco los resultados esperados. De ser así, se creyó riesgoso implementar un piloto con el modelo como herramienta, que finalmente pudiera terminar siendo perjudicial para la organización.

Teniendo en cuenta lo anterior, se decidió utilizar Machine Learning con foco en conocer más a los repartidores que se habían inactivado, y un modelo que permitiera clasificarlos en función de su probabilidad de reactivación, para luego poder proponer un piloto experimental que corrroborara las inferencias presentadas en función de los resultados obtenidos, junto con permitir mejorar la venta de rutas en Wareclouds. Se optó por esto ya que con los datos disponibles, este uso de Machine Learning podría tener mejor desempeño, así como también, teniendo en cuenta que no había algún estudio previo de los clouders, se consideró adecuado comenzar conociéndolos mejor, junto con tener la posibilidad de disminuir los riesgo con la propuesta experimental.

Bibliografía

[1] Random oversampling and undersampling for imbalanced classification, 2020. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/.

[2] Smote for imbalanced classification with python, 2020. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[3] Wareclouds: el Airbnb de la logística y de las ventas online, 2021. https://www.latercera.com/pulso/noticia/wareclouds-el-airbnb-de-la-logistica-y-de-las-ventas-online/5ELYRXPVLJBWZBNSHJPUTZWBPQ/.

[4] Wareclouds, la startup chilena de logística colaborativa, iniciará operaciones en México, 2021. https://contxto.com/es/chile/wareclouds-startup-logistica-colaborativa-mexico/.

[5] Growth Wareclouds, 2022. https://datastudio.google.com/u/0/reporting/bfc79bb7-7845-4cb3-ba1e-ff00fbf34ca1/page/6zXD.

[6] Sector centro de Santiago, 2022. https://es.wikipedia.org/wiki/Santiago_(comuna).

[7] Sector nororiente de Santiago, 2022. https://es.wikipedia.org/wiki/Sector_nororiente_de_Santiago.

[8] Sector norponiente de Santiago, 2022. https://es.wikipedia.org/wiki/Sector_norponiente_de_Santiago.

[9] Sector norte de Santiago, 2022. https://es.wikipedia.org/wiki/Sector_norte_de_Santiago.

[10] Sector sur de Santiago, 2022. https://es.wikipedia.org/wiki/Sector_sur_de_Santiago.

[11] Sector suroriente de Santiago, 2022. https://es.wikipedia.org/wiki/Sector_suroriente_de_Santiago.

[12] Sector surponiente de Santiago, 2022. https://es.wikipedia.org/wiki/Sector_surponiente_de_Santiago.

[13] Cristopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[14] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[15] Pete Chapman, Julian Clinton, and Randy Kerber. *CRISP-DM 1.0 Step by step guide*. 2000.

[16] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. *Journal of artificial intelligence research*. 16:231-357, 2002.

[17] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras TensorFlow*. O'reilly Assoc Inc, 2019.

Anexos

Anexo A

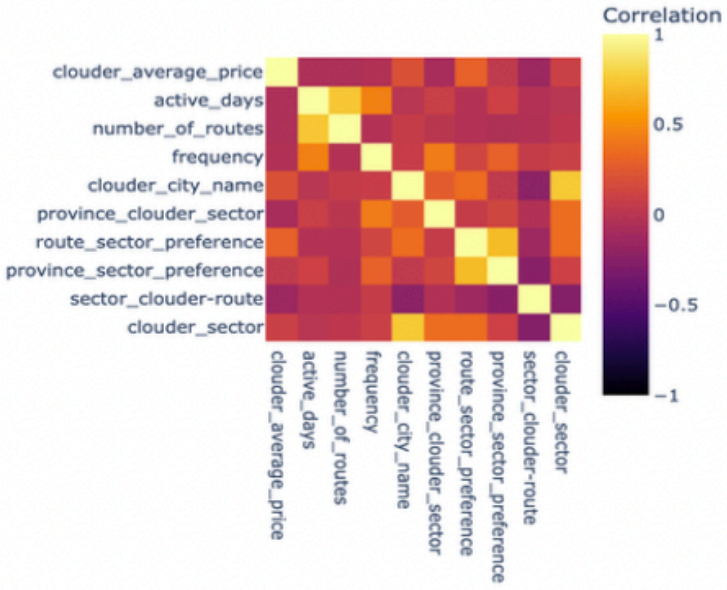

Figura A.1: Matriz de correlación de las variables de los clouders

Anexo B

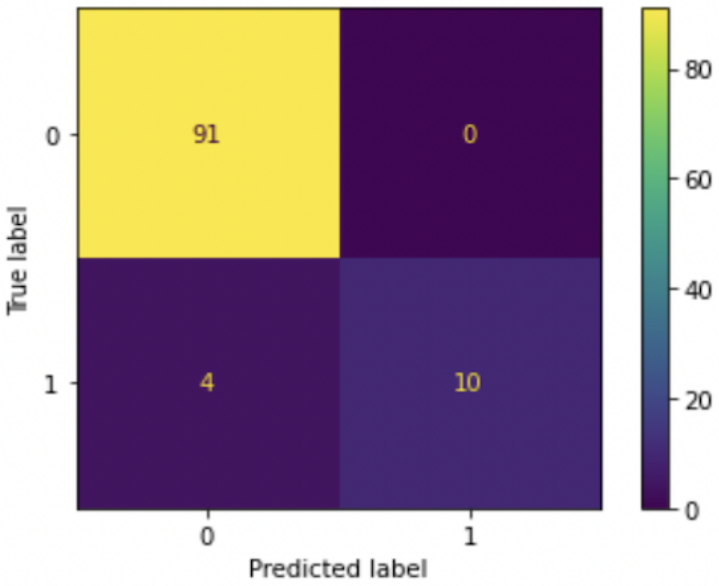

Figura B.1: Matriz de confusión de algoritmo Árbol de decisión con conjunto de entrenamiento

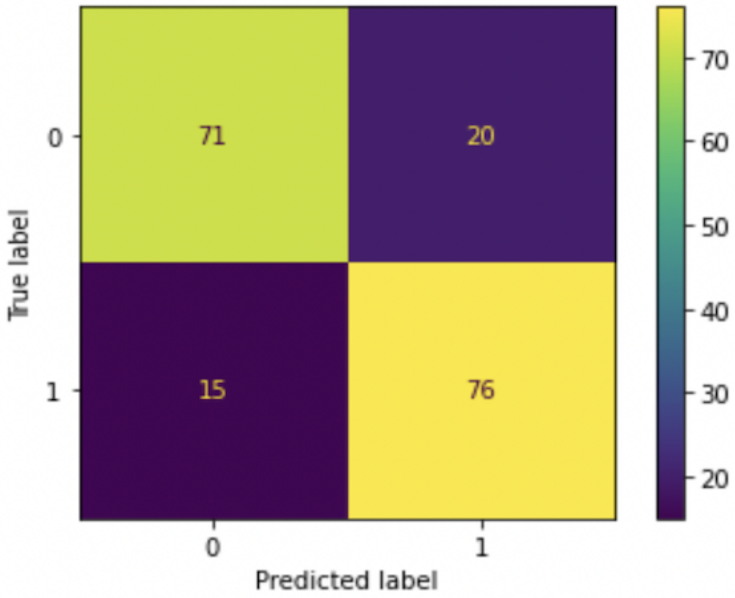

Figura B.2: Matriz de confusión de algoritmo Regresión logística con conjunto de entrenamiento

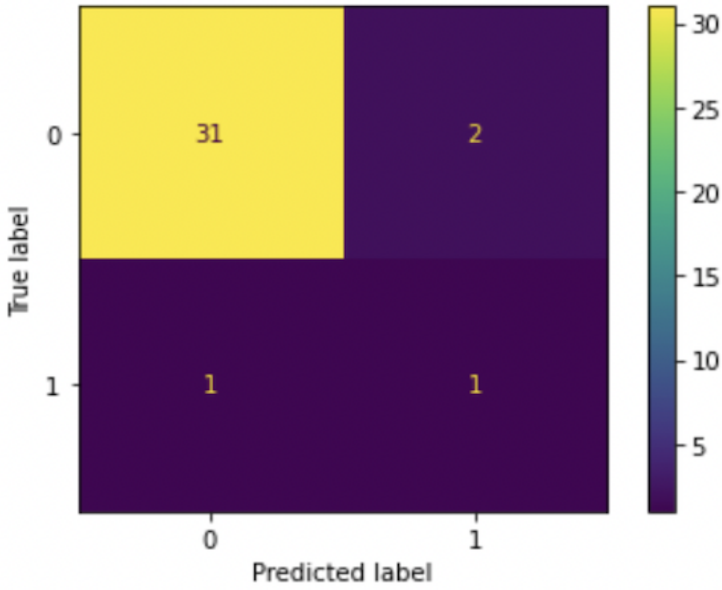

Figura B.3: Matriz de confusión de algoritmo Árbol de decisión con conjunto de test

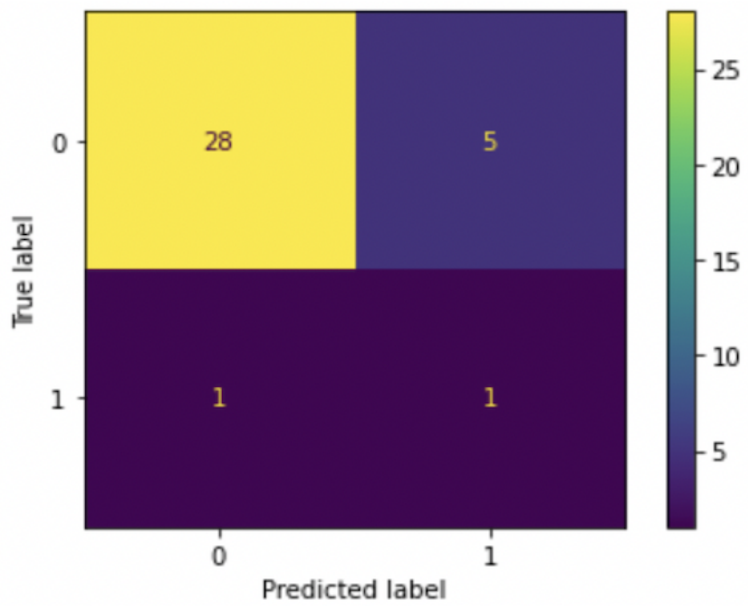

Figura B.4: Matriz de confusión de algoritmo Regresión logística con conjunto de test

Anexo C

C.1. Escenario optimista

	Rutas mensuales que deben ser vendidas manualmente	Proyección de rutas mensuales que deberán ser vendidas manualmente	Proyección rutas sin vender	Personas necesarias (s/ proyecto)	Costo empresa personas extra (s/ proyecto)	Rutas que deban ser vendidas manualmente	Rutas que deban ser vendidas manualmente	Rutas sin vender (c/proyecto)	Personas necesarias (c/ proyecto)	Personas extras (c/proyecto)	Costo empresa personas (c/ proyecto)
Enero	80										
Febrero	80										
Marzo	80										
Abril		80	4	1	820000	-592	0	0	1	0	820000
Mayo		100	5	2	1640000	-572	0	0	1	0	820000
Junio		280	14	4	3.280.000	-392	0	0	1	0	820000
Julio		360	18	5	4.100.000	-312	0	0	1	0	820000
Agosto		480	24	6	4.920.000	-192	0	0	1	0	820000
Septiembre		620	31	8	6.560.000	-52	0	0	1	0	820000
				TOTAL	**21.320.000**					**TOTAL**	**4920000**
Rutas asignadas automáticamente (por mes)	320									**AHORRO**	**16.400.000**
Rutas vendidas manual (por mes)	80									**%**	**77**
Clouders activos	40										
Promedio de rutas tomadas por cada clouder (por mes)	8										
Clouders reactivados (escenario optimista)	84										
N° de rutas que tomarían clouders reactivados	672										
Personas necesarias hoy	1										
Costo empresa trabajador	820000										
% rutas no vendidas de las rutas manual	0,05										

Figura C.1: Cálculo del ahorro en el costo de contratar nuevas personas dedicada a la venta de rutas manual

	Rutas mensuales que deben ser vendidas manualmente	Proyección de rutas mensuales que deberán ser vendidas manualmente	Proyección rutas sin vender	Rutas vender manual con clouders reactivados (optimista)	Rutas sin vender (c/proyecto)	% fuga (s/ proyecto)	Marcas fugadas (s/proyecto)	Costo fuga (s/ proyecto)	% fuga (c/ proyecto)	Marcas fugadas (c/proyecto)	Costo fuga (c/ proyecto)	Proyección marcas por mes
Enero	80											
Febrero	80											
Marzo	80											
Abril		80	4	0	0	0,02	1	650000	0	0	0	224
Mayo		100	5	0	0	0,03	9	5850000	0	0	0	281
Junio		280	14	0	0	0,07	26	16900000	0	0	0	365
Julio		360	18	0	0	0,09	43	27950000	0	0	0	474
Agosto		480	24	0	0	0,12	74	48100000	0	0	0	616
Septiembre		620	31	0	0	0,16	129	83850000	0	0	0	801
							282	183300000		TOTAL	0	
N° marcas fugadas por despachos atrasados	3									AHORRO	183300000	
% fuga	0,02									%	100	
Costo fuga promedio por marca	650000											
% rutas no vendidas de las rutas manual	0,05											

Figura C.2: Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega

	Proyección pedidos
Abril	10914
Mayo	13643
Junio	17.736
Julio	23.056
Agosto	29.973
Septiembre	38.964
Total pedidos (6 meses)	134.286
Total pedidos (1 año)	268.572
Promedio pedidos mensuales	22.381
Promedio pedidos diarios	1120
Aumento en cada evento (60%)	1792
N° de rutas total	72
N° de rutas vendidas manualmente (s/ proyecto)	29
N° marcas fugadas (s/ proyecto)	3
Costo fuga por marca	1100000
N° de eventos al año	5
Costo fuga s/ proyecto por evento	3300000
Costo fuga anual (s/proyecto)	16500000
Clouders nuevos	84
N° de rutas vendidas manualmente (c/ proyecto)	0
N° marcas fugadas (c/ proyecto)	0
Costo fuga c/ proyecto por evento	0
Costo fuga anual (c/proyecto)	0
AHORRO	16500000
%	100

Figura C.3: Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega en días de *peaks* de pedidos

C.2. Escenario medio

	Rutas mensuales que deben ser vendidas manualmente	Proyección de rutas mensuales que deberán ser vendidas manualmente	Proyección rutas sin vender	Personas necesarias (s/ proyecto)	Costo empresa personas extra (s/ proyecto)	Rutas que deban ser vendidas manualmente	Rutas que deban ser vendidas manualmente	Rutas sin vender (c/proyecto)	Personas necesarias (c/ proyecto)	Personas extras (c/proyecto)	Costo empresa personas extra (c/ proyecto)
Enero	80										
Febrero	80										
Marzo	80										
Abril		80	4	1	820000	-312	0	0	1	0	820000
Mayo		100	5	2	1640000	-292	0	0	1	0	820000
Junio		280	14	4	3.280.000	-112	0	0	1	0	820000
Julio		360	18	5	4.100.000	-32	0	0	1	0	820000
Agosto		480	24	6	4.920.000	88	88	0	2	1	1640000
Septiembre		620	31	8	6.560.000	228	228	0	3	2	2460000
				TOTAL	**21.320.000**					**TOTAL**	**7380000**
Rutas asignadas automáticamente (por mes)	320									**AHORRO**	**13.940.000**
Rutas vendidas manual (por mes)	80									**%**	**65**
Clouders activos	40										
Promedio de rutas tomadas por cada clouder (por mes)	8										
Clouders reactivados (escenario medio)	49										
N° de rutas que tomarían clouders reactivados	392										
Personas necesarias hoy	1										
Costo empresa trabajador	820000										
% rutas no vendidas de las rutas manual	0,05										

Figura C.4: Cálculo del ahorro en el costo de contratar nuevas personas dedicada a la venta de rutas manual

	Rutas mensuales que deben ser vendidas manualmente	Proyección de rutas mensuales que deberán ser vendidas manualmente	Proyección rutas sin vender	Rutas vender manual con clouders reactivados (medio)	Rutas sin vender (c/proyecto)	% fuga (s/ proyecto)	Marcas fugadas (s/proyecto)	Costo fuga (s/ proyecto)	% fuga (c/ proyecto)	Marcas fugadas (c/proyecto)	Costo fuga (c/ proyecto)	Proyección marcas por mes
Enero	80											
Febrero	80											
Marzo	80											
Abril		80	4	0	0	0,02	1	650000	0	0	0	224
Mayo		100	5	0	0	0,03	9	5850000	0	0	0	281
Junio		280	14	0	0	0,07	26	16900000	0	0	0	365
Julio		360	18	0	0	0,09	43	27950000	0	0	0	474
Agosto		480	24	88	5	0,12	74	48100000	0,03	19	12350000	616
Septiembre		620	31	228	12	0,16	129	83850000	0,06	49	31850000	801
							282	183300000		TOTAL	44200000	
N° marcas fugadas por despachos atrasados	3	0,02								AHORRO	139100000	
% fuga	0,02									%	75,89	
Costo fuga promedio por marca	650000											
% rutas no vendidas de las rutas manual	0,05											

Figura C.5: Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega

	Proyección pedidos
Abril	10914
Mayo	13643
Junio	17.736
Julio	23.056
Agosto	29.973
Septiembre	38.964
Total pedidos (6 meses)	134.286
Total pedidos (1 año)	268.572
Promedio pedidos mensuales	22.381
Promedio pedidos diarios	1120
Aumento en cada evento (60%)	1792
N° de rutas total	72
N° de rutas vendidas manualmente (s/ proyecto)	29
N° marcas fugadas (s/ proyecto)	3
Costo fuga por marca	1100000
N° de eventos al año	5
Costo fuga s/ proyecto por evento	3300000
Costo fuga anual (s/proyecto)	16500000
Clouders nuevos	49
N° de rutas vendidas manualmente (c/ proyecto)	0
N° marcas fugadas (c/ proyecto)	0
Costo fuga c/ proyecto por evento	0
Costo fuga anual (c/proyecto)	0
AHORRO	16500000
%	100

Figura C.6: Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega en días de *peaks* de pedidos

C.3. Escenario pesimista

	Rutas mensuales que deben ser vendidas manualmente	Proyección de rutas mensuales que deberán ser vendidas manualmente	Rutas sin venderProyección rutas sin vender	Personas necesarias (s/ proyecto)	Costo empresa personas extra (s/ proyecto)	Rutas que deban ser vendidas manualmente	Rutas que deban ser vendidas manualmente	Rutas sin vender (c/proyecto)	Personas necesarias (c/ proyecto)	Personas extras (c/proyecto)	Costo empresa personas extra (c/ proyecto)
Enero	80										
Febrero	80										
Marzo	80										
Abril		80	4	1	820000	-32	0	0	1	0	820000
Mayo		100	5	2	1640000	-12	0	0	1	0	820000
Junio		280	14	4	3.280.000	168	168	0	3	2	2460000
Julio		360	18	5	4.100.000	248	248	0	4	3	3280000
Agosto		480	24	6	4.920.000	368	368	0	5	4	4100000
Septiembre		620	31	8	6.560.000	508	508	0	7	6	5740000
				TOTAL	**21.320.000**					**TOTAL**	**17220000**
Rutas asignadas automáticamente (por mes)	320									**AHORRO**	**4.100.000**
Rutas vendidas manual (por mes)	80									**%**	**19**
Clouders activos	40										
Promedio de rutas tomadas por cada clouder (por mes)	8										
Clouders reactivados (escenario pesimista)	14										
N° de rutas que tomarían clouders reactivados	112										
Personas necesarias hoy	1										
Costo empresa trabajador	820000										
% rutas no vendidas de las rutas manual	0,05										

Figura C.7: Cálculo del ahorro en el costo de contratar nuevas personas dedicada a la venta de rutas manual

	Rutas mensuales que deben ser vendidas manualmente	Proyección de rutas mensuales que deberán ser vendidas manualmente	Proyección rutas sin vender	Rutas vender manual con clouders reactivados (pesimista)	Rutas sin vender (c/proyecto)	% fuga (s/ proyecto)	Marcas fugadas (s/proyecto)	Costo fuga (s/ proyecto)	% fuga (c/ proyecto)	Marcas fugadas (c/proyecto)	Costo fuga (c/ proyecto)	Proyección marcas por mes
Enero	80											
Febrero	80											
Marzo	80											
Abril		80	4	0	0	0,02	1	650000	0	0	0	224
Mayo		100	5	0	0	0,03	9	5850000	0	0	0	281
Junio		280	14	168	9	0,07	26	16900000	0,05	19	12350000	365
Julio		360	18	248	13	0,09	43	27950000	0,07	34	22100000	474
Agosto		480	24	368	19	0,12	74	48100000	0,1	62	40300000	616
Septiembre		620	31	508	26	0,16	129	83850000	0,13	105	68250000	801
							282	183300000		TOTAL	143000000	
N° marcas fugadas por despachos atrasados	3									AHORRO	40300000	
% fuga	0,02									%	21,99	
Costo fuga promedio por marca	650000											
% rutas no vendidas de las rutas manual	0,05											

Figura C.8: Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega

	Proyección pedidos
Abril	10914
Mayo	13643
Junio	17.736
Julio	23.056
Agosto	29.973
Septiembre	38.964
Total pedidos (6 meses)	134.286
Total pedidos (1 año)	268.572
Promedio pedidos mensuales	22.381
Promedio pedidos diarios	1120
Aumento en cada evento (60%)	1792
N° de rutas total	72
N° de rutas vendidas manualmente (s/ proyecto)	29
N° marcas fugadas (s/ proyecto)	3
Costo fuga por marca	1100000
N° de eventos al año	5
Costo fuga s/ proyecto por evento	3300000
Costo fuga anual (s/proyecto)	16500000
Clouders nuevos	14
N° de rutas vendidas manualmente (c/ proyecto)	15
N° marcas fugadas (c/ proyecto)	2
Costo fuga c/ proyecto por evento	2200000
Costo fuga anual (c/proyecto)	11000000
AHORRO	5500000
%	33,33

Figura C.9: Cálculo del ahorro en el costo de fuga de marcas por incumplimiento en los plazos de entrega en días de *peaks* de pedidos