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

DESARROLLO DE UN MODELO DE PREDICCIÓN DE FUGA DE CLIENTES Y DISEÑO DE EXPERIMENTO PARA LA APLICACIÓN DE ESTRATEGIAS DE FIDELIZACIÓN EN FACTORING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JAVIER ANDRÉS SEGURA MORENO

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
CÉSAR ARAYA HERNÁNDEZ
MARÍA FERNANDA VARGAS COURBIS

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: INGENIERO CIVIL INDUSTRIAL
POR: JAVIER ANDRÉS SEGURA MORENO
FECHA: 2022
PROF. GUÍA: PABLO MARÍN VICUÑA

DESARROLLO DE UN MODELO DE PREDICCIÓN DE FUGA DE CLIENTES Y DISEÑO DE EXPERIMENTO PARA LA APLICACIÓN DE ESTRATEGIAS DE FIDELIZACIÓN EN FACTORING

El presente trabajo de título se desarrollará en la empresa Chita Spa. Esta empresa pertenece al rubro del factoring, el cual cosiste en prestar un servicio financiero anticipando dinero a otras empresas (negocio B2B) a cambio de la cesión de una factura con una deuda asociada. El foco de clientes de Chita son las PYMEs del mercado chileno.

En base a la metodología CRISP-DM, se logró comprender en profundidad el negocio donde el principal hallazgo se relaciona con la tasa de fuga, problema latente para la Gerencia Comercial. Para esto, se realizó una búsqueda exhaustiva de los datos disponibles dentro de la empresa para explicar el comportamiento de los clientes y calcular los niveles de fuga. Se evidencia que desde el 2018 hasta agosto de 2021, en promedio, hay una tasa de fuga mensual de un 37 %. En base a este problema, se construyó una base de datos la cual alimentó el proceso iterativo de modelamiento, donde se probó, evaluó y eligió un modelo de clasificación del tipo Árbol de Clasificación, obteniendo un rendimiento general de un 73.6 % en la correcta clasificación de clientes activos y fugados. Donde las variables más importantes a la hora de etiquetar son la *Recencia*, la *Antigüedad*, el *Monto* acumulado de las facturas operadas, los *Días de mora* y la *Hora de giro* hacia el cliente.

Con la elección de este modelo, se continuó con la elaboración de una propuesta de medidas comerciales para incentivar la reoperación de los clientes potenciales a fugarse. Ofreciendo un descuento de un 20 % para la tasa de interés, 15 % de descuento para la comisión y la fijación del porcentaje de anticipo en 98 %, que son costos de servicio directos que el cliente percibe.

Posteriormente, se formuló un diseño experimental para medir el impacto de identificar proactivamente (con los modelos de clasificación) a clientes que puedan fugarse, de manera que se definió un total teórico de 550 clientes requeridos para poder medir un 7.6 % del tamaño del efecto en la tasa de fuga entre los grupos de control y tratamiento, los que se alcanzan en 5 meses de operación. Finalmente, se realizó una simulación financiera del experimento (corto plazo) y una simulación de implementación de la estrategia (mediano plazo), comparando las utilidades y cantidad de clientes acumulados para diferentes niveles de fuga y retención. De esto se obtuvo que a corto plazo la inversión hace que las utilidades por hacer el experimento sean menores (-10.7 %), mientras que los clientes aumentan (+6.1 %). Por su parte y complementando estos resultados, con la simulación a 12 meses se obtuvo que tanto las utilidades como los clientes acumulados aumentan en un 3.3 % ($28.545.380) y un 7.2$ (34), respectivamente. De manera que su eventual implementación, en base a las simulaciones realizadas, a mediano plazo Chita recupera su inversión.

Agradecimientos

A mi mamá, Bárbara, Rodrigo, Viviana y como no olvidar a mi papá que donde esté me está apoyando. Gracias por estar al lado mío desde pequeño, incentivando a que siempre me superara y llegara lo más lejos que pudiera, dándome las herramientas y mucho amor. Gracias por darme ejemplo qué imitar y qué no imitar para siempre ser una mejor persona. Gracias por el apoyo y amor incondicional. Gracias a mis sobrinos hermosos que siempre me dan la energía de seguir por ser su ejemplo.

A todas las personas que permitieron que hiciera esta memoria. Gracias Cristián Murúa, por siempre tenerme fe y darme tus tips de trabajador con experiencia. Por pelearla por mi cuando había que conseguir algo y por confiar en mi trabajo siempre. A mis profesores, Pablo Marín y César Araya. A pesar de a veces sentir que los feedback's eran duros, muchas gracias por ser parte de este proceso.

A mi amigo Christian que no se cansó nunca de preguntarme cómo iba con la memoria, que como el hermano de otra familia siempre creyó en mi. Gracias por cultivar tanta cosa buena desde que casi nacimos. Al Fele que de un minuto a otro se volvió un amigo más, compañero de noches y juego, gracias por siempre creer en mi y darme ánimo hasta más no poder. A mis amigos del colegio, Ary, Basty, Nacho, Yogo, Agurto, Tobal y Jair (te considero uno más), por tanto apañe y apoyo siempre. Por tenerme siempre una fe altísima, por brindarme tantos buenos momentos, apoyo, alegría, amor y contención. A la Cami y el Agus que fueron un pilar para mi en buenos, malos y terribles momentos dentro de la universidad. Gracias Cami por ser mi pana desde que estuvimos en inducción, pasando por muchas locuras, nuestros términos de carrera y mucho más, de verdad gracias por todo y por siempre creer en mi. Gracias Agus por de repente llegar a mi vida darme apoyo, buenos momentos, harto carrete y una amistad bacán. Gracias por siempre tenerme fe en todo.

A la Marce, la Negra, por estar y ser desde el minuto 1 mi fan. Gracias por apañarme siempre y complementarnos como el gran team que somos en todo lo que implica la universidad y la vida misma. Gracias por brindarme tanta sabiduría y por guiarme cuando mi cabeza se nublaba y no podía salir del hoyo en el que estaba. Gracias por llenarme de tanto amor, tanto aprecio, tanto carrete, tanto viaje y tanta energía que se traspasó a las páginas de esta memoria. Gracias por creer en mi cuando ni yo creía en mi y por hacer de mi una persona más noble y un mejor ingeniero. De verdad, gracias!

Finalmente, quiero agradecer a todas las personas que en algún momento pasaron por mi vida universitaria, a compañeros, compañeras, profesores, profesoras, auxiliares, ayudantes, auxiliares de aseo, etc, que fueron parte de mi formación universitaria y que hicieron de alguna manera que sea la persona soy. Muchas gracias a todos y todas... Se logró !

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Características de la empresa

Chita es una fintech [1] creada en la segunda mitad del año 2016 y está ubicada en la ciudad de Santiago de Chile. Al año 2021, brinda servicios financieros a empresas del tipo Business to Business (B2B) exclusivamente a través de su página web. Para esto utiliza una metodología de trabajo que está basada en tecnología y con la que se busca innovar en un rubro de larga data y requerido a nivel mundial.

1.1.1. Rubro

Desde la década de los 80 aproximadamente se instauró en Chile una forma de dar liquidez a empresas que necesitaban un apoyo financiero de corto plazo para solventar necesidades básicas de sus negocios como las remuneraciones, compras de insumos o alguna otra necesidad a cubrir. A esta forma de financiamiento entre empresas se le denominó Factoring o también conocido como Factoraje. El servicio de factoring permite dar liquidez mayoritariamente a micro, pequeñas o medianas empresas a través de un anticipo de dinero.

Dado que es un sector económico que presenta interacciones solamente entre empresas, este actúa como un negocio B2B. De esta manera, Chita se posiciona en el mercado financiero chileno como una fintech de factoring.

1.1.2. Misión y visión

La entidad financiera Chita declara su misión como:
"Queremos darle acceso a financiamiento a todas las PYMEs de Chile, de forma Simple, Rápida y para Todos".

Y por su lado, declara su visión como:
"Queremos ser la mejor alternativa de financiamiento de las PYMEs en LATAM"

1.1.3. Organigrama

Como se mencionó previamente, Chita es una empresa financiera que fue creada en el año 2016. Al comienzo tenía una estructura organizacional simple, o sea solo con jefaturas principales. Si bien esta estructura se mantuvo con una tendencia a la horizontalidad (sin tantas

jerarquías) durante un largo periodo, con el paso del tiempo ha crecido, se ha complejizado y agregado o modificado algunos niveles como gerencias, jefaturas nuevas o cargos intermedios. Esos cambios se han ido incorporando dada la necesidad de contar con una estructura organizacional acorde al crecimiento operacional desarrollado. El resultado final de toda esa evolución estructural se ve reflejado en la Figura 1.1.

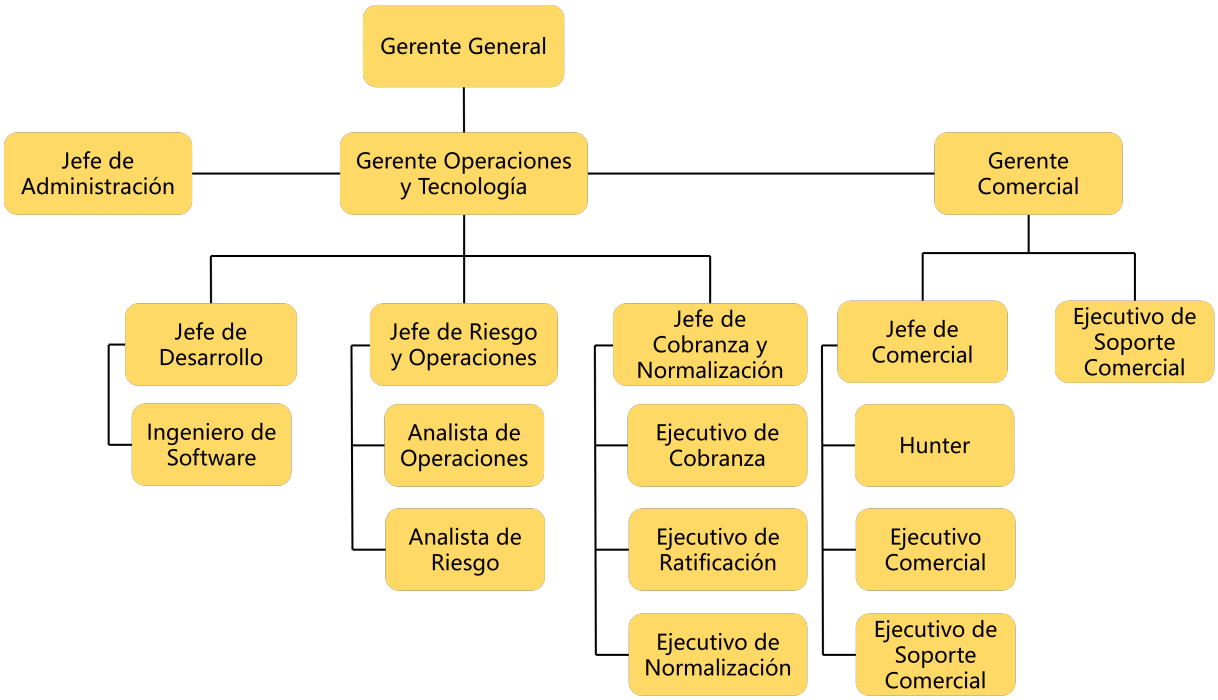


Figura 1.1: Organigrama de Chita.

1.1.4. Servicios

Desde la creación de Chita, ésta solo ha prestado el servicio de factoring, lo que se ha extendido hasta el 2021 concentrando todos los esfuerzos en ese único servicio.

El factoring tiene varias aristas y modalidades en su implementación. Este servicio, como herramienta financiera, se define como un contrato que se celebra entre dos partes, pero que en su totalidad involucra a tres partes. Mediante este contrato, una empresa traspasa o cede las facturas que ha emitido y a cambio obtiene un anticipo de dinero, según lo define la Comisión para el Mercado Financiero, desde ahora CMF [2]. En específico, Chita utiliza el tipo de factoring de créditos por venta de facturas. Para que esta definición sea más rigurosa, cada una de estas tres partes tienen una denominación específica, según su función dentro del proceso.

Las tres partes involucradas en la operación son (definidas desde el punto de vista de la empresa financiera):

- Empresa de factoring:
  Parte que anticipa un monto de dinero al cliente (en general no el 100 % para tener un excedente en caso de un retraso en el pago del deudor).

- Cliente:
  Parte (mayoritariamente Pymes) que solicita anticipo de dinero respectivo a las facturas cedidas a la empresa de factoring.

- Deudor:
  Parte que cancela la deuda que tiene con el cliente a través de facturas emitidas por el pago de un servicio o producto y que con este proceso termina cancelando la deuda a la empresa de factoring.

Entre estas tres partes se genera un flujo de acciones que desencadena la realización adecuada del proceso, según las funciones anteriormente descritas. En primer lugar, el cliente debe contactar a la empresa de factoring para iniciar el proceso. Durante esta etapa, la entidad financiera solicita información del cliente y las facturas a ceder para poder hacer la evaluación de riesgo pertinente y definir el porcentaje final de anticipo. Si se acepta el riesgo, se llevan a cabo todas las acciones legales correspondientes para hacer efectiva la cesión de la factura y se le proporciona el dinero acordado al comienzo al cliente. Dentro del contrato, están estipuladas las condiciones y plazos en que el deudor debe cancelar el monto de la factura que ahora le corresponde saldar con la empresa de factoring. Si el deudor cancela dentro de lo acordado, se cierra el proceso satisfactoriamente. En caso contrario existen dos etapas. Primero se contacta al cliente para que tome conocimiento de la situación de no pago y tome acciones con el deudor, pues es el cliente quien debe asumir la responsabilidad legal y financiera del contrato; y su vez se contacta al deudor para que pague el dinero adeudado. Si esta situación persiste durante un periodo estipulado en el contrato, tal como se mencionó, se toman acciones legales contra el cliente, quien debe asumir todo el pago que el deudor no realizó. Todo este proceso se puede ver reflejado de manera resumida en la Figura 1.2.

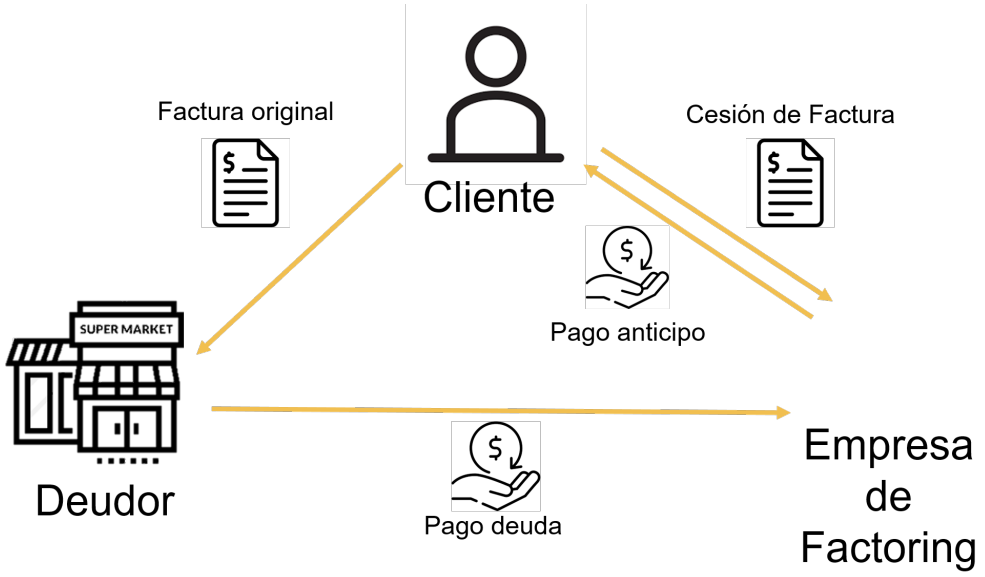


Figura 1.2: Diagrama del Factoring.

1.1.5. Clientes y actividad realizada

Debido al crecimiento económico del país, muchos emprendimientos o empresas se han ido conformando y consolidando poco a poco en el mercado chileno. Sin embargo, todo inicio

de una empresa es un proceso complejo visto desde múltiples dimensiones, en específico la relacionada con el financiamiento. La cual es uno de los pilares fundamentales, sino el más importante, para poder crear un negocio y hacer que este prospere a lo largo del tiempo. En particular, las empresas que más preocupadas están del financiamiento son aquellas que tienen tamaños acotados y con capacidades no tan robustas para conseguir capital y para sustentar los distintos procesos internos.

Es por esta razón que Chita, dado su rol financiero dentro del mercado de empresas chilenas, se enfoca mayoritariamente en conseguir clientes que, según las definiciones de tamaño según ventas del SII, van desde el primer rango microempresa hasta el segundo rango mediana empresa [3]. En otras palabras, tiene un enfoque a pequeñas, micro y medianas empresa (PYMEs). A pesar de establecer su foco en organizaciones con ese tamaño, de igual forma ofrece sus servicios a la totalidad de los rangos. Pero como se mencionó anteriormente, tiene un enfoque predominante a empresas de menor tamaño, centrándose más en la cantidad de clientes de esas características que en el volumen del monto de las facturas operadas.

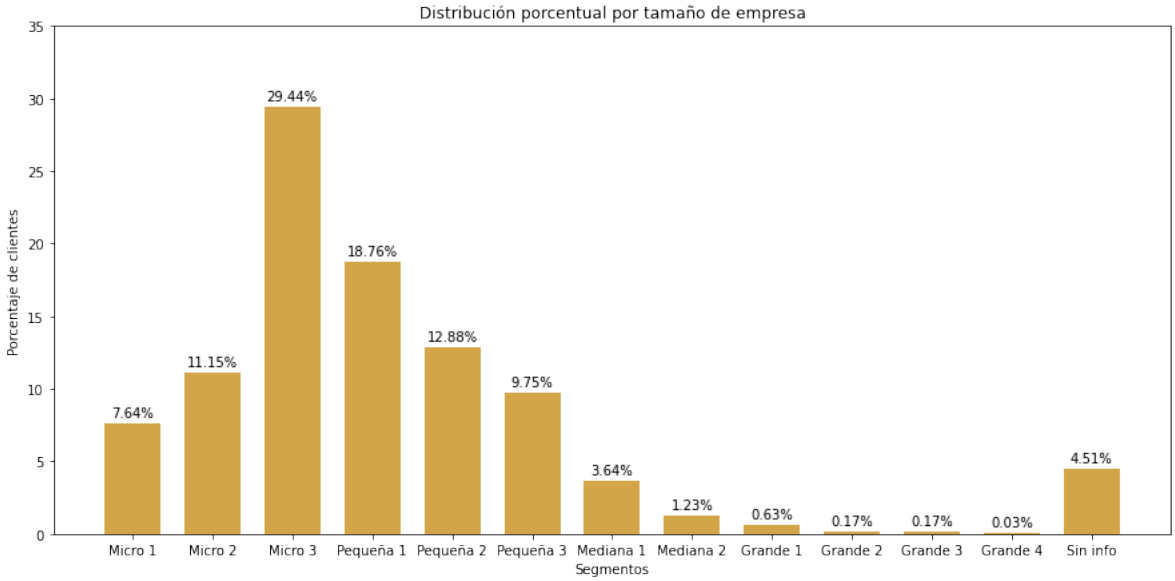


Figura 1.3: Gráfico de distribución por tamaño de clientes operados por Chita (Elaboración propia).

Al año 2021, Chita ha operado históricamente con alrededor de 3.000 clientes de distintos tamaños y rubros. Concentrando aproximadamente un 95 % de su cartera histórica entre micro, pequeñas y medianas empresas. Por otro lado, concentra al 91 % de sus clientes entre los siete rubros que más participación han tenido, tal como se puede ver en la Figura 1.4.

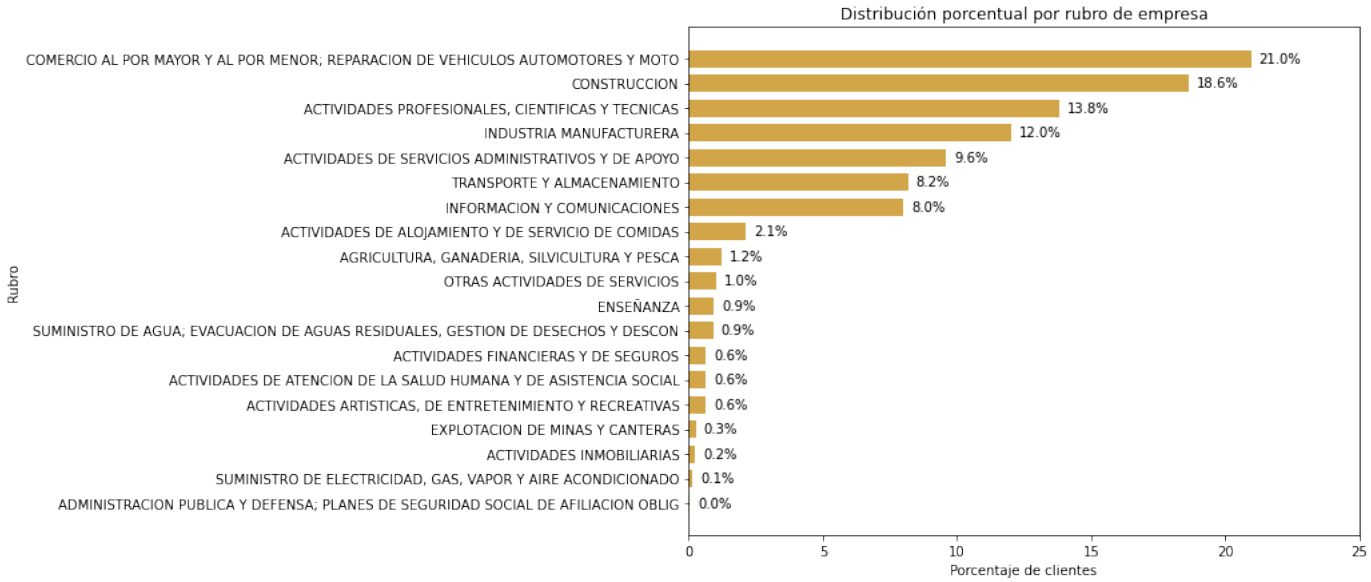


Figura 1.4: Gráfico de distribución de rubros operados por Chita (Elaboración propia).

Adicionalmente, al ser una fintech, Chita opera de manera cien por ciento remota, es decir, que no tiene sucursales o puntos de interacción con clientes. Dicho de otra forma, la empresa no conoce a sus clientes de manera presencial, sino que solo a través de canales digitales o telefónicos.

1.2. Mercado

1.2.1. Niveles de operación

Desde fines del año 2016 Chita está presente en el sector del factoring, teniendo cuatro años completos de operación (2017 – 2020). En particular, para la industria descrita a lo largo del presente informe, las empresas definen una métrica operacional para medir cuánto dinero financiaron a sus clientes (anticipos) llamada colocaciones. Esta métrica no representa directamente ganancia o ingresos, pero mide el flujo de dinero que pasa por la empresa y logra prestar, a través de su servicio, a sus clientes. En vista de que Chita solo tiene cuatro años completos de operación, se reportan tres crecimientos de sus colocaciones en el mercado. Para el año 2018 alcanzó un crecimiento, respecto al año anterior, de un 141 %. A su vez, para el año 2019 un 76 % y, finalmente, para el año 2020 un -13 %.

Por otra parte, si se analiza la operación considerando instituciones de factoring bancario como no bancario, es decir, el mercado completo, se tiene que las colocaciones llegan a un total aproximado equivalente a un 8,3 % del Producto Interno Bruto (PIB) de Chile [4][5]. Ahora bien, Chita en comparación con su mercado, financia alrededor de un 0,2 % del total de las colocaciones.

1.2.2. Marco regulatorio

En general, para el sector económico o financiero existen muchas regulaciones desde entidades como la Superintendencia de Valores y Seguros (SVS) o la evolución actual de la

misma, integrada con la Superintendencia de Bancos e Instituciones Financieras (SBIF) que es la CMF. Sin embargo, en particular para el rubro del factoring una las principales regulaciones vienen desde la ley N°18.045 [6], la cual en su primer artículo declara (en resumen) que se regularán las entidades financieras que inviertan en la bolsa de valores, sean sociedades abiertas o sean emisores de oferta pública. Debido a esta ley y a que Chita fue parte de otra entidad financiera que tenía inversiones en la bolsa, cumplió con regulaciones desde la CMF. Sin embargo, hoy en día que es una empresa independiente y no cumple con las condiciones que la ley indica, no es regulada por la CMF. No obstante, a través de sus proveedores legales que ayudan a generar contratos para las operaciones con las facturas y a auditar este proceso, se cumplen las normas para validar todos los instrumentos financieros involucrados como lo piden las instituciones del estado.

1.2.3. Tendencias y posicionamiento de mercado

Si bien no existe una gran desagregación y cantidad material disponible público para analizar a la industria del factoring adecuadamente como ocurre con otras, se puede observar, según cifras generales de asociaciones de factoring, que este mercado viene desde hace por lo menos cinco a seis años en un constante crecimiento, tanto a nivel de colocaciones como a nivel de empresas que lo conforman. Al año 2021, existen entre 150 a 200 empresas que participan de este sector industrial, financiando a aproximadamente un 5 % de las empresas en el mercado chileno, según estudios internos de la organización.

Dentro de las empresas que ejercen y prestan el servicio del factoring existen dos clasificaciones, el bancario y no bancario. El primero maneja colocaciones a un nivel muy alto en término de los montos de las facturas a diferencia del no bancario (operaciones con grandes empresas generalmente). Dado este contexto, se podría establecer que Chita es una empresa seguidora a nivel general del mercado, pues pertenece a la clasificación de factoring no bancario. Sin embargo, se puede hacer otra distinción dentro del factoring, ya que hay empresas como la analizada que son entidades financieras tecnológicas (fintech). Dentro de esta categoría, en base a las ventajas competitivas y al nivel de colocaciones que tiene Chita, se podría establecer que está bien posicionada en ese submercado.

Es importante mencionar que las tendencias han tenido fuertes cambios debido al contexto que se vive mundialmente respecto a la situación sanitaria del COVID-19. Si se mira al número de empresas desde el 2020 hasta 5 años atrás, se puede ver que la tendencia era claramente al alza. Sin embargo, el cierre anual desde 2019 a 2020 tuvo una caída de aproximadamente 30.000 (-2,3 %) empresas, por lo que el mercado de las empresas se vio enfrentado a forzosos cierres (Ver Figura 1.5).

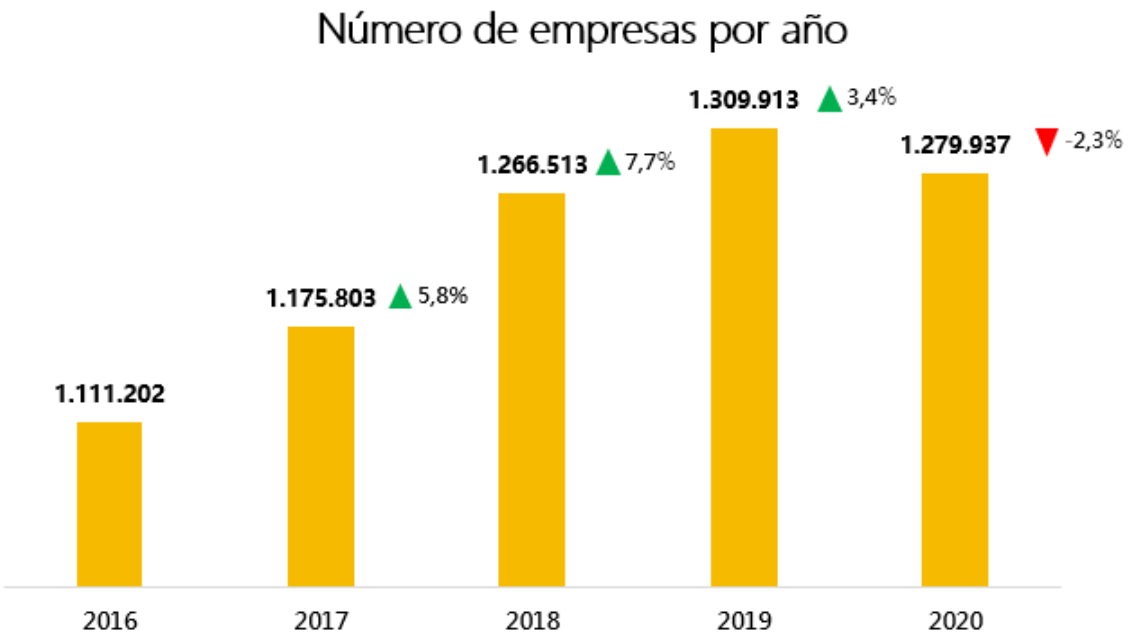


Figura 1.5: Gráfico que muestra la evolución del número de empresas en Chile desde el 2016 hasta el 2020 [7].

1.3. Desempeño organizacional

En las secciones anteriores se presentaron los crecimientos para los años en operación y existencia que lleva la organización analizada. Obteniendo crecimientos de 141 % para 2018, 76 % para 2019 y, por último, -13 % para 2020. Como se aprecia, desde la creación de la empresa se tuvo un sostenido crecimiento durante dos años. Desafortunadamente, el contexto sanitario del COVID-19 afectó los mercados financieros, en especial a los segmentos objetivo que se tienen (PYMEs) pues son los más sensibles a las restricciones aplicadas. Si bien, en un comienzo de la pandemia, como protección, muchas empresas solicitaron financiamiento a través del factoring, esto tuvo una fuerte caída los siguientes meses. Lo cual terminó repercutiendo en las colocaciones que tuvo Chita, haciendo que tuviera un decrecimiento para el año 2020 de un 13 %. A pesar de tener un crecimiento menor (2020) a los primeros años, según los antecedentes que tiene la empresa en análisis, se puede decir que la empresa está en etapa de crecimiento en relación a su progreso a lo largo de los años de operación.

Capítulo 2

Descripción del proyecto y justificación

2.1. Principales funciones, actores y relación interna

La Gerencia Comercial, dentro de la organización, es el motor del negocio, debido a que lo central en el factoring no bancario es conseguir gran volumen de clientes, pues el margen por cliente es bajo[1]. Es por esto que el Área Comercial es la encargada de la generación y captación de clientes nuevos y gestión de clientes antiguos para poder hacer que operen nuevamente. Además, es la encargada de la gestión de agentes, para poder referir nuevos clientes y capturar nuevas alianzas para este proceso. De esta forma, para poder entender el rol de las funciones, se profundizará cada una a continuación.

- Captación nuevos clientes
  Chita tiene un enfoque, a diferencia de empresas de factoring grandes como bancos, en la cantidad de clientes que capta por sobre los montos de las propias facturas que se operan. En otras palabras, se tiene un foco en volumen de clientes más que otra variable. Por tanto, es muy relevante tener una constante gestión para conseguir nuevos prospectos de clientes (a cargo del equipo de Hunters) los que luego se convierten en clientes de la empresa.

- Reoperación de clientes
  Una vez que un cliente opera, pasa a una categoría de cliente antiguo. Al igual que para la función anterior, existe un equipo especializado (ejecutivos/as comerciales) y enfocado en clientes que ya operaron con la empresa, donde el objetivo principal es acercarse a esos clientes y gestionar una nueva transacción. De manera que se pueda fidelizar a los clientes y, en caso de que el cliente no responda a esos estímulos, retenerlos con medidas comerciales más agresivas.

- Programa de agentes
  Este programa, es una forma alternativa al marketing digital para conseguir clientes nuevos. Su funcionamiento se implementa a dos niveles, para personas naturales y para

[1] El margen por cliente bajo se debe a que en general el rubro del factoring se caracteriza por ser un mercado de bajo margen debido a que las ganancias se deben capturar a nivel de tasas de interés y comisiones. Adicionalmente al foco de la empresa en estudio es ser competitiva ya que la cartera de clientes se compone sobre un 95 % de Pymes.

empresas. La lógica del programa es que cualquiera de los niveles anteriores refiera a un cliente nuevo, recibiendo de forma única una comisión una vez que ese cliente referido opere con Chita. Así la organización tiene otra fuente externa a su personal de trabajo para captar clientes nuevos. Además, existe una variante del programa, en la cual un agente[2] puede tener un nivel de agentes bajo su tutela. Sin embargo, tiene una fuerte restricción de que se le otorgan comisiones por referencias de los agentes tutelados solo para un nivel hacia abajo.

- Alianzas antiguas y nuevas
  Todas estas funciones, en algunos puntos de la gestión, se ven apoyadas por alianzas que se tienen con otras organizaciones. Por ejemplo, se tienen alianzas con empresas que son facturadores electrónicos o ayudan con plataformas ERP's que permiten tener una mejor gestión de personas y conexión interna entre Hunters, ejecutivos/as comerciales, agentes, etc. Por esto, es muy importante mantener las alianzas ya construidas y estar en una constante búsqueda de nuevas alianzas que puedan ayudar a mejorar y optimizar la operación del negocio.

Con estas funciones descritas, se tiene una idea más acabada de las preocupaciones y las acciones que la Gerencia Comercial debe realizar para poder generar un adecuado funcionamiento de la organización de cara a los clientes. Como se mencionó previamente, esta área es importante dentro de la empresa respecto a la captación y gestión de clientes. Por lo tanto, tener un buen desempeño en término de los objetivos planteados como estrategia, tiene por consecuencia un mayor flujo de clientes y posible crecimiento.

Por otra parte, respecto al tamaño de la Gerencia Comercial, esta se compone de diez personas en total. Ahora bien, su estructura es bastante simple, tiene dos ejes principales, uno que se encarga de la gestión de clientes y otro de la gestión de alianzas. Para la gestión de clientes, se tiene a la jefatura comercial y bajo esta, ejecutivos/as comerciales, hunters y un soporte comercial especialmente con preocupación en el programa de agentes. Y desde el otro eje, se tiene solo un soporte comercial encargado de gestionar alianzas antiguas para mantenerlas y conseguir nuevas que puedan a ayudar a la operación. Esto se puede apreciar de una forma más clara en el diagrama mostrado en la Figura 2.1.

[2] Persona o empresa que tiene convenio con Chita en programa de agentes y es quien refiere a clientes nuevos.

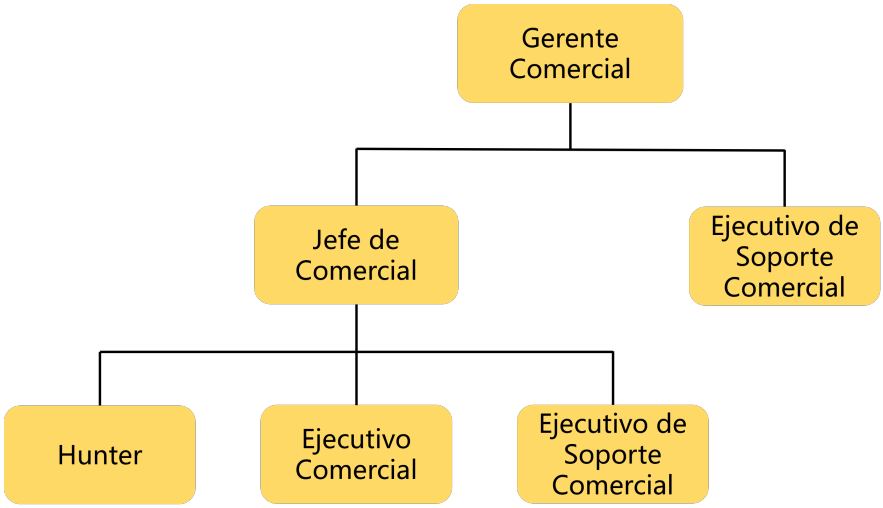


Figura 2.1: Organigrama Gerencia General de Chita (Elaboración propia).

Tal como se puede distinguir en el organigrama general de la empresa y según las funciones de la Gerencia Comercial, ésta tiene relación con el resto de las áreas pero no tiene clientes directos, sino que el Área Comercial es cliente de las demás partes de la organización. Es decir, la Gerencia en estudio se centra en la recepción de insumos internos como reportes de riesgo, operacionales y del Área de Tecnologías de Información.

2.2. Identificar el problema u oportunidad

Así como se pudo apreciar en el apartado anterior, el trabajo de título de desarrollará en el área que genera el movimiento principal de la operación de toda la organización, en términos del flujo de clientes, ya sea para adquirir nuevos, gestionar antiguos y recuperar o incentivar a quienes no han vuelto a operar en un tiempo definido.

Chita es una empresa que se enfoca en el segmento de las PYMEs, las cuales según los datos transaccionales del año 2020 operaron un monto promedio de aproximadamente $1.6MM y una mediana de $0.6MM. Es por esto que la organización se centra, más que en conseguir facturas de montos altos, en captar un gran volumen de clientes. Debido a esta definición de estrategia comercial de negocio, es muy importante monitorear constantemente indicadores de adquisición y fuga de clientes. Pues, como el interés primordial es la cantidad de clientes en sus distintos estados, se debe considerar la siguiente regla de negocios: **la tasa de clientes nuevos o adquiridos debe ser mayor o igual a la de fugados en un periodo de tiempo establecido**. Si esta condición no se cumple la cantidad de clientes activos sería menor a la de clientes inactivos, aumentando a posibilidad de generar pérdidas.

En vista de que se han mencionado los distintos tipos de clientes, resulta necesario definir estas tres clasificaciones.

- Cliente nuevo: es aquel usuario que en un determinado mes comienza por primera vez un proceso de operación de facturas a través del servicio de factoring y la empresa de factoring le gira el dinero (conversión).
- Cliente activo: es aquel cliente que ya operó al menos una vez con la empresa, sigue

operando o se mantiene sin realizar transacciones por menos de 60 días desde la última operación.

- Cliente fugado: es aquel cliente activo que permanece sin operar por 60 días o más. Sin embargo, este tipo de cliente tiene dos categorías. Cliente inactivo, cuando se mantiene entre 60 y 179 días sin operar. Y Cliente Perdido, cuando se mantiene más de 180 días sin operar.

De esta manera es intuitivo revisar las tasas de clientes nuevos y de clientes fugados. No obstante, antes de revisar las tendencias mencionadas, es importante entender cómo se comportan los clientes en relación a los tiempos entre operaciones. Se debe tener en consideración que para este análisis se dividió al total de clientes en dos grupos, quienes operaron solo una vez con la empresa durante el año 2020 y quienes operaron dos o más veces durante el mismo periodo. Puesto que hay clientes que son operadores de emergencia (necesidad del momento) o clientes que prefirieron a la competencia.

Figura 2.2: Gráfico que muestra la tendencia de la cantidad de clientes únicos y tiempos entre operaciones promedio para cada año (clientes con al menos 2 operaciones).

Como se puede ver en la Figura 2.2, la cantidad de clientes únicos operados desde que la empresa inició sus operaciones ha crecido constantemente hasta el inicio de la pandemia (2020), donde en un principio hubo una gran demanda pues muchas PYMEs querían protegerse del incierto escenario económico que podría venir. Sin embargo, al paso de uno o dos meses de iniciada la pandemia la demanda bajó drásticamente (se puede observar este fenómeno con mayor detalle en la sección de análisis exploratorio), lo que tuvo como consecuencia en que la empresa en estudio no creció en cantidad de clientes, ni tampoco en colocaciones (ver Anexo A.1). Dado el contexto de los datos es posible comparar desde el año 2017 al 2020, pues para el año 2021 solo se consideran 8 meses de datos, lo que representa dos tercios de los meses utilizados para los otros meses. La tendencia muestra que hubo un crecimiento en clientes por sobre el 100 % hasta el año 2020 donde se esperaba seguir al alza, lo cual no ocurrió reduciéndose la cantidad de clientes en un 5 %.

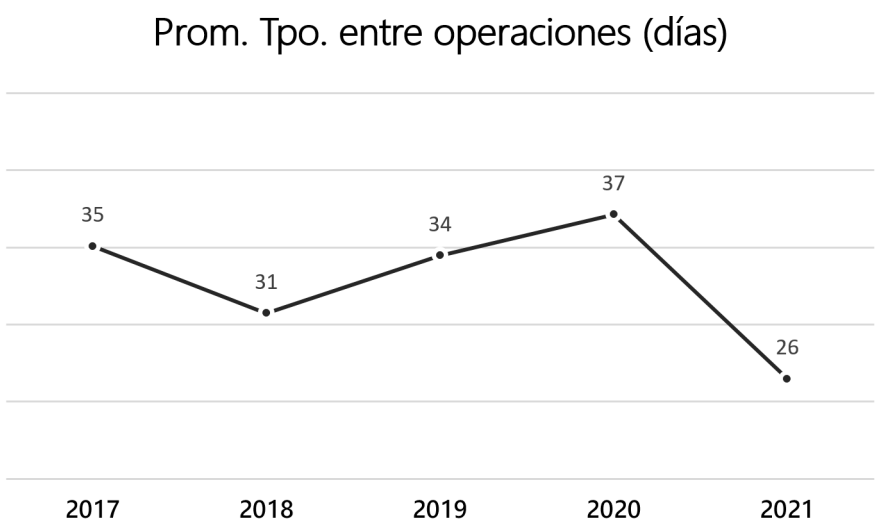


Figura 2.3: Gráfico que muestra la tendencia de los tiempos entre operaciones promedio para cada año (clientes con al menos 2 operaciones).

Por otro lado, se tiene el análisis de los tiempos entre operaciones que se puede ver en la Figura 2.3 Si bien, parece mantenerse relativamente constante (en promedio) y sin considerar al año 2021 que está afectado por el desbalance de datos respecto a los otros años, se puede notar que este valor venía a la baja en los primeros años mientras que el volumen de clientes crecía. Sin embargo, métricas como el tiempo entre operaciones se vio perjudicado por factores externos relacionados al mercado, aumentando su valor en relación a valores de años previos (desde 2019 en adelante). Si bien los tiempos promedio se ven no tan altos en relación al umbral de un cliente inactivo (60 días), se puede ver a continuación la distribución de los clientes que operaron el año 2020.

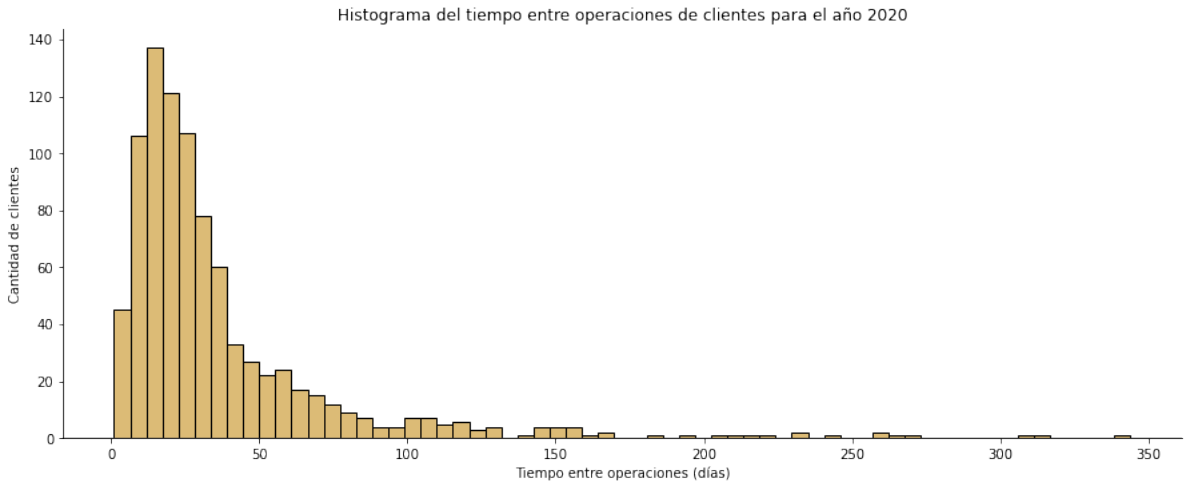


Figura 2.4: Histograma que muestra distribución de los tiempos entre operaciones de los clientes para el año 2020 (clientes con al menos 2 operaciones).

Tal como se observa, a pesar de que el promedio son 37 días entre operaciones, hay 134 (15 %) clientes que tienen un promedio mayor a 60 días entre sus operaciones lo cual es riesgoso, puesto que algunos cambian de estado antes de volver a operar y que en el peor de

los casos terminan perdiéndose y no volviendo a operar nuevamente con la organización en estudio.

Aunque es relevante entender los intervalos de tiempo que transcurren entre las operaciones de los clientes y la tendencia del volumen de clientes a lo largo de los años, es importante para el caso de estudio analizar la tasa de clientes nuevos y tasa de fuga. Para realizar el análisis de estas tasas se debe tener en cuenta que no es tan simple realizar una comparación entre ellas, pues la tasa de clientes nuevos se puede analizar en ventanas acotadas de tiempo. Por el contrario, la tasa de fuga debe analizarse en un periodo de tiempo más extenso. Es por este motivo que el primer acercamiento hacia la tasa de fuga se realiza a nivel global, en término de los años de operación. Para este cálculo se revisan las tasas para los tres últimos años completos de operación, o sea, 2018, 2019 y 2020 (ver Figura 2.5).

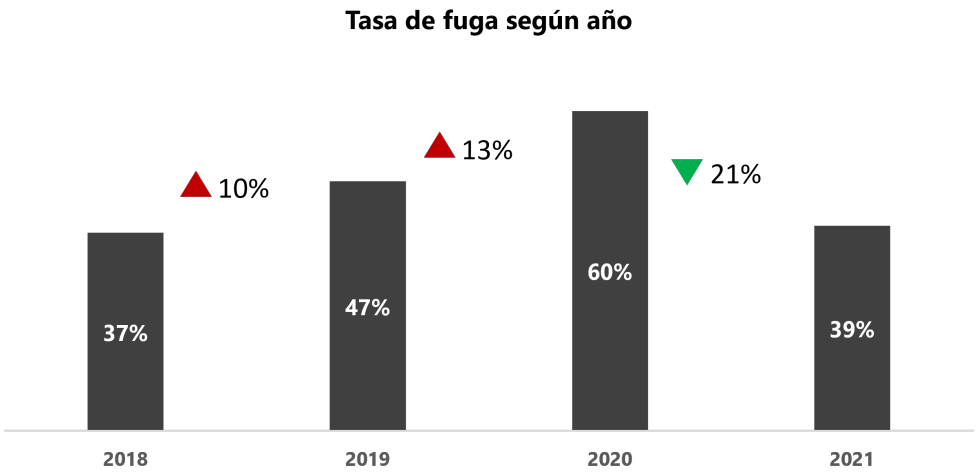


Figura 2.5: Gráfico que muestra la tasa de fuga (clientes con al menos 2 operaciones).

Al analizar la tasa de fuga, se puede notar que los valores no son despreciables, pues los valores mostrados están cerca de una tasa del 50 % lo que representa que al final de año, de los clientes que operaron, solo la mitad ha operado en menos de 60 días hasta el último día del año. Más aún cuando el foco es tener un gran volumen de clientes recurrentes, es decir que los clientes que operan puedan seguir haciéndolo durante el tiempo. Para este análisis se consideraron exclusivamente clientes que operaron como mínimo dos veces en los años de estudio, dado que es son grupo poco comparables a nivel anual (más adelante se analizan conjuntamente). Como se ve en la Figura 10, la tasa de fuga desde el 2018 tuvo una sostenida alza hasta finales del tercer periodo analizado, aumentando un 10 % y 13 % en los años 2019 y 2020, respectivamente. Esto último muestra que para el año 2020 hubo un peak de tasa de fuga alcanzando un 60 %, lo cual se condice con un efecto de la pandemia Covid-19 (ver detalle mensual en sección de análisis exploratorio). Si bien no es el único factor, es relevante el movimiento que causó sobre las empresas activas en Chile (Ver Figura 5), donde solo entre marzo y abril de 2020 hubo 75 mil empresas que dejaron de estar activas. En particular, dado que el foco de la empresa en estudio son la Pymes, dejaron de estar activas 17 mil pequeñas empresas y mil medianas empresas.

Por lo tanto, es en esta temática donde se encuentra un problema a resolver y una gran

oportunidad para lograr mejorar las estrategias de identificación de potenciales clientes de fuga, estrategias de fidelización o retención lo cual se debería verse reflejado en las métricas de fuga y la gestión de clientes. Ya que como se planteó, tener cada vez una menor tasa de fuga de clientes implica que el volumen de clientes recurrentes aumente y se mantengan operando durante un periodo más largo de tiempo. Además, según se vio para los tiempos entre operaciones, a medida que se tienen más clientes operando (efecto de reducir la tasa de fuga) menor podrían ser los tiempos entre cada transacción, aumentando la frecuencia de los clientes y, por lo tanto, los ingresos percibidos.

Los detonantes de que la tasa de fuga tenga los niveles mostrados podrían estar relacionados a diferentes factores que hasta el momento no se están observando por la empresa ni tampoco en este análisis macro que se ha realizado. En primer lugar, pueden estar afectando factores de experiencia como lo son posibles problemas al ceder facturas, el tiempo que pasa desde la cesión de una factura hasta el giro del anticipo, entre otras variables que se incluirán dentro de estudio. Otro factor puede ser la competencia de otras empresas de factoring, sin embargo, este estudio se centrará en la operación interna que se tiene con los clientes, por lo que la competencia no se considerará.

Causas importantes para relevar a nivel interno, es que Chita desde el Área Comercial, está respondiendo a destiempo con los clientes que no han tenido interacciones con la empresa en un tiempo más acotado que lo definido para un cliente fugado. Es decir, se podría estar teniendo una respuesta muy reactiva y tardía frente a la inactividad de un cliente. De modo que se tendrá un enfoque en la anticipación o proactividad de las acciones comerciales.

Capítulo 3

Objetivos

3.1. Objetivo General

El objetivo general del trabajo de memoria consiste en: "Identificar proactivamente a clientes activos potenciales de fuga para aumentar el volumen de clientes recurrentes, utilizando modelos de predicción que sirvan como base para aplicar estrategias comerciales y diseños experimentales".

3.2. Objetivos Específicos

Para poder cumplir con lo propuesto en el objetivo general, se establecen los siguientes objetivos específicos:

- Identificar las variables relevantes a nivel de negocio (comportamiento y experiencia del cliente) para predecir la fuga de clientes, seleccionarlas e incluirlas en el modelo de predicción de fuga.
- Construir, evaluar y seleccionar el mejor modelo de predicción de fuga, obteniendo una herramienta reutilizable para etiquetar clientes activos potenciales de fuga.
- Construir, en base al modelo de predicción, una propuesta de estrategia comercial de retención (paquetes de descuentos) y diseño experimental para su futura implementación y medición.
- Desarrollar propuesta y plan de implementación[3] del proyecto, para dejar bien informado y definido como se debe implementar a futuro este proceso, de modo que sea un insumo útil y sostenible en el tiempo.

3.3. Alcances y Resultados esperados

3.3.1. Alcances

Tanto la elaboración de los modelos de predicción de fuga como la construcción de la propuesta de estrategias comerciales y diseño experimental respectivo, se enmarcará exclusivamente dentro de las funciones de la Gerencia Comercial definidas anteriormente. Esto se

[3] Documento que detallará características técnicas, pasos a seguir y consideraciones a tener para la implementación real en el funcionamiento de la empresa del trabajo de memoria.

corresponde con la utilización de bases de datos que abarcan información transaccional de la empresa (operación de facturas), las cuales tienen consigo información del cliente y de riesgo. Adicionalmente, se usarán bases de datos que contienen información de captura de leads como también data específica de los rendimientos de las campañas de marketing digital. Como fuente externa a la empresa en estudio se agregó información obtenida desde el Servicio de Impuestos Internos (SII), la cual contiene datos generales y demográficos de las empresas que han iniciado actividades, de modo que se pueda cruzar con las bases de datos internas y obtener un conjunto de datos más completo.

Asimismo, la temporalidad de los datos utilizados como input de los modelos abarcará transacciones realizadas desde el 2017 hasta el mes de septiembre de 2021. Como se mencionó en el documento, la empresa inició sus actividades operacionales a fines del año 2016, por lo tanto, existe muy poca información para analizar de aquel año y es valioso hacer comparaciones de métricas anuales, lo cual no se podría llevar a cabo. A pesar de tener información desde el año 2017, se utilizarán datos desde el mes de enero de 2018 hasta septiembre de 2021 para aplicar en los modelos de machine learning, debido a que los datos del año 2017 presentan inconsistencias a nivel transaccional que pueden provocar ruido al momento de procesar la información.

Si bien, dentro del desarrollo del presente trabajo se construye una propuesta a nivel comercial para aplicar en base a los resultados de los modelos, no se podrá implementar puesto que el tiempo de desarrollo del trabajo de título es acotado comparado al tiempo necesario para implementar las etapas de la propuesta. Es por esta misma razón que se comentó y expuso dentro de los objetivos específicos, que se elaborará un plan de implementación (mencionado en la sección de objetivos) con el detalle de cómo realizar el proceso de forma independiente y sin ayuda del memorista, entregándolo una vez terminado el trabajo de título. De modo que, si en un futuro Chita quiere implementarlo autónomamente, lo pueda realizar sin contratiempos. Lo que significa que la implementación final del proyecto quedará en manos del solicitante, o sea del Gerente Comercial.

3.3.2. Resultados esperados

En base a los objetivos y metodología planteada en el desarrollo de este documento, se espera conseguir una construcción adecuada de un modelo de predicción de fuga que pueda identificar de manera eficaz los clientes activos que tienen una mayor propensión a la fuga según su comportamiento histórico y experiencia. Para lograr esto, es necesario balancear su rendimiento en base a distintas métricas que se deben relacionar a nivel comercial con una implementación futura, teniendo en cuenta la cantidad de clientes y los costos asociados.

Además, se espera poder construir una propuesta robusta para aplicar medidas comerciales de retención y su propia medición a través de un diseño experimental. El objetivo es poder crear un paquete (no muy extenso por las características del negocio) de ofertas o descuentos que impacten directamente en los costos que los clientes deben incurrir para usar el servicio que la empresa en estudio presta, o sea en las tasas de interés o comisiones cobradas. Por parte del diseño experimental, se espera poder dejar una propuesta en base a la cantidad de clientes del último año móvil, de modo que se deje planteado cuántos clientes se necesitan para que el proceso de medición pueda tener validez estadística y a su vez planificar el tiempo de duración del experimento. Así, se podría verificar empíricamente el impacto del proyecto

desarrollado por el memorista.

Finalmente, se espera poder construir un completo plan de implementación, donde quede correctamente documentado el proceso, cómo se utilizan las herramientas desarrolladas durante el trabajo de título y cómo hacer que estas se puedan extender y generalizar a un mediano y largo plazo. Permitiendo a la empresa y, en particular, al solicitante tener herramientas que posibiliten tener una mejor respuesta ante posibles fugas de clientes que lo que se tiene actualmente.

Capítulo 4

Marco Conceptual

4.1. Fuga de clientes

De acuerdo con lo mencionado acerca del problema de estudio, se quiere desarrollar una herramienta que permita enfrentar la fuga de clientes. Sin embargo, para entender en profundidad lo que eso significa, es necesario tener una definición del concepto de fuga. Según [8], "un fugado es un cliente que dejó la compañía". Si bien esta definición es general y sin tanto detalle, en la sección 2.2 del presente trabajo se tiene una definición más precisa respecto al caso que se está tratando (cliente activo, inactivo y perdido). Para efectos de este proyecto, el concepto de "dejar la compañía" está directamente relacionado con la operación del cliente y el tiempo que pasa desde su última operación con la empresa. En particular, para el desarrollo de los modelos se considerará a los clientes fugados inactivos, que son clientes activos que tienen entre 60 y 180 días sin volver a operar. Se toma esta decisión debido a que los clientes fugados perdidos no pertenecen a una clasificación cercana al cliente activo y el objetivo central es anticiparse y poder evitar su entrada en la categoría de inactivo.

4.2. Modelos de predicción

4.2.1. Regresión Logística

La Regresión Logística es uno de los modelos más utilizados dentro del machine learning para predicción de variables categóricas y probabilidades, especialmente cuando se habla del estudio de fuga de clientes. Como dice su nombre es una regresión, por ende, se compone por una variable dependiente (y) que será de carácter binario y de variables explicativas (x). Sin embargo, para este tipo de modelo se incluye además una componente probabilística, la cual está definida según una distribución de probabilidad que está asociada al comportamiento del fenómeno que se quiere estudiar.

Para esto, se define la probabilidad de ocurrencia de un evento como se muestra en la ecuación 4.1 [9][10]:

$$P(y = 1|x) = f(z) = \frac{1}{(1 + e^{-z})} \tag{4.1}$$

Donde,

$$z = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \tag{4.2}$$

Con k variables explicativas $(x_1, \ldots, x_k)$ y $k+1$ coeficientes o parámetros $(\beta_0, \beta_1, \ldots, \beta_k)$.

Una vez que ya se tiene la base de la regresión logística, hay que establecer cuál será la metodología utilizada dentro del modelo para conseguir el objetivo, o sea, ajustar los parámetros utilizados en la ecuación 4.2. Para lograr esto, se utilizará el método de *máxima verosimilitud (ML)*, el cual, a diferencia de los mínimos cuadrados, considera interacciones no lineales dentro de su construcción y desarrollo, tal como se necesita para un modelo de regresión logística. Un ejemplo de esto es que el método ML no requiere que las variables explicativas o independientes posean restricciones dentro del proceso de optimización. Así, se define a LL como la función de log-verosimilitud [9], la cual para fines de entrenar el modelo se debe maximizar según se muestra en la ecuación 4.3.

$$LL(\beta) = \sum_{i=1}^{k} y_i log(f(z_i)) + (1 - y_i) log(1 - f(z_i)) \tag{4.3}$$

Este tipo de regresiones tiene una similitud muy grande con los modelos de regresión lineal básicos, lo que a través de los parámetros de la función z resulta más fácil de interpretar, para sí tener una mejor noción de los resultados y su relación con las preguntas de negocio que se quieren responder. Además, dada la simplicidad que presenta respecto a su construcción versus otros modelos que requieren más de una distribución de probabilidad para desarrollarse o formas funcionales que necesitan mucho más procesamiento computacional para implementarse, la regresión logística es uno de los modelamientos más eficientes y que menos recursos consume a la hora de ejecutarse en un computador. Por otro lado, dada las características de los datos de la empresa (pocos años de historia de datos y una cantidad reducida de clientes), este modelo es una opción factible y adaptable a lo que el trabajo de título demanda.

Una de las desventajas más importantes a considerar para este tipo de modelamiento es la sensibilidad que se tiene respecto a valores anómalos o outliers que las bases de datos contengan. Es decir, que los resultados de la regresión logística tienden a sesgarse rápidamente cuando los datos que recibe como input presenta valores que están muy fuera de la distribución de la variable explicativa a la que pertenece. No obstante, hay metodologías que modifican la función objetivo, incluyendo otras formas funcionales que provocan que el desempeño del modelo sea bastante más robusto frente a outliers, sin tener la necesidad de eliminarlos o dejarlos fuera del análisis [11].

4.2.2. Árbol de Clasificación

Los Árboles de Decisión son modelos que pueden ser usados para dos fines: Clasificación (denominados *Árboles de Clasificación*) y Regresión (denominados *Árboles de Regresión*). Además, son muy utilizados por la simplicidad e interpretabilidad que estos brindan al momento de relacionar los resultados con la hipótesis o pregunta de negocio. De modo que resulta mucho más intuitivo de analizar e incluirlo en la toma de decisiones sobre algún fenómeno estudiado.

Para el caso de estudio de este trabajo de título se quiere clasificar a los clientes dentro de

una categoría binaria (cliente activo o cliente potencial de fuga), por tanto, la denominación que se utilizará de este momento en adelante para el modelo será de Árbol de Clasificación.

Los Árboles de Clasificación tienen una lógica fácil de comprender. A medida que el algoritmo va avanzando en su ejecución, va dividiendo a los datos según reglas transparentes y de rápido entendimiento, resultando grupos de datos más homogéneos que la división original (raíz del árbol) en términos del resultado de la variable dependiente. Con el fin de realizar este proceso, se define a la variable dependiente (y), la cual debe ser categórica, específicamente binaria para este desarrollo. Además, se definen k variables explicativas o independientes ($x_1, \ldots, x_k$). Como se mencionó anteriormente, la lógica del algoritmo [12] es particionar recursivamente el espacio de trabajo en k dimensiones (el mismo número de variables explicativas) rectangulares que no se pueden sobreponer.

Esta división sigue un proceso iterativo que inicia eligiendo una variable independiente x_i la cual tiene un valor s_i, usándose para hacer la primera división del espacio. Concretamente, el algoritmo toma el valor s_i y lo utiliza como pivote, es decir, genera dos conjuntos de datos. Un grupo que cumple con la condición $x_i \leq s_i$ y, por su parte, el otro grupo debe seguir la desigualdad que complementa a la del primero grupo, o sea, $x_i > s_i$. De esta forma, se itera con el resto de las $k-1$ variables explicativas alcanzando lo propuesto en un principio, dividir el espacio en k grupos. Considerando que con cada división las agrupaciones ya hechas, supone una mayor pureza. Definiendo a un conjunto puro como aquel que solo contiene una clasificación de un solo tipo. Si bien ese es el objetivo, no siempre se consigue y al terminar las k divisiones siguen existiendo dos hojas en los nodos finales del árbol.

A pesar de haber expuesto la lógica del algoritmo para dividir el espacio según las variables predictivas, se debe definir el criterio con el que se va eligiendo la variable explicativa que divide al espacio. Con el propósito de tomar esa decisión, se recurre dos métricas que buscan medir la pureza de los grupos resultantes de la división. Estas son la Entropía e Impureza de Gini y se definen en las ecuaciones 4.4 y 4.5, respectivamente.

Dado un espacio A y m clases de la variable dependiente definido por $c = 1, \ldots, m$.

$$Entropía(A) = -\sum_{c=1}^{m} p_c log_2(p_c) \tag{4.4}$$

$$I(A) = 1 - \sum_{c=1}^{m} p_c^2 \tag{4.5}$$

Donde p_c representa la proporción de observaciones en el espacio A pertenecientes a la clase c. Para estas métricas hay que considerar que valores de p iguales a cero significa que las observaciones son iguales dentro del conjunto y $1-1/m$ cuando los elementos tienen igual proporción. Con todo esto definido, finalmente se llega a la elección de variable explicativa para la división, la que dependerá exclusivamente de encontrar la máxima diferencia entre la impureza del nodo $(i-1)$ y el nodo i. Dicho en otras palabras, la variable que divide a grupos lo más homogéneos posible. Así, iterativamente el algoritmo entrena al modelo para poder, posteriormente, usarlo para clasificar datos nuevos.

Los Árboles de Clasificación no presentan reglas matemáticas muy complejas y tienen una

fácil interpretabilidad. Asimismo, es un modelo que permite ser graficado y visualizar todas las decisiones que se toman para llegar a los resultados (no es una "caja negra"), siempre y cuando el número de niveles que tenga el árbol lo permita. Por otro lado, es robusto en el sentido de tratar con outliers y trabaja muy bien con variables categóricas, a diferencia de otros modelos de machine learning. Sin embargo, una de las desventajas es que para entrenarlo y conseguir buenos resultados se necesita una cantidad no despreciable de observaciones en las bases de datos que se tengan como input.

4.2.3. Evaluación de los modelos

Debido a que se tendrán dos tipos de modelo a evaluar, es importante definir, en líneas generales, las métricas que se utilizarán para poder discriminar qué algoritmo tiene un mejor desempeño según el objetivo propuesto: *identificar a clientes activos potenciales de fuga y a clientes activos*.

Primero, es necesario hacer una separación de la base de datos con la que se entrenará el modelo de predicción de fuga. Para esto se dividirá la base de datos en una proporción de 80-20 % correspondiente a entrenamiento y testeo, respectivamente. El propósito de hacer este paso intermedio antes de evaluar los modelos es poder utilizar una porción de la información para que los modelos aprendan (80 %). Luego, testear con información que nunca haya sido introducida a los modelos (20 %) revisando su rendimiento frente a esos nuevos datos, de modo que los modelos no aprendan de memoria la información que reciben de input y puedan tener un real poder de predicción.

Una vez que el proceso anteriormente descrito se haya realizado, se deben comparar los modelos en término de los valores de las métricas de rendimiento. Como el problema a resolver en el trabajo de título es de clasificación binaria (activo o fugado), una de las formas para evaluar el desempeño de un modelo es mirando una tabla llamada *Matriz de Confusión*. Como lo dice su nombre, es una matriz que muestra el cruce de las etiquetas que un cierto modelo da como resultado o predice. Donde las filas de la matriz representan los valores reales de los datos, mientras que las columnas muestran los valores que predice el modelo [13] (Ver Figura 4.1).

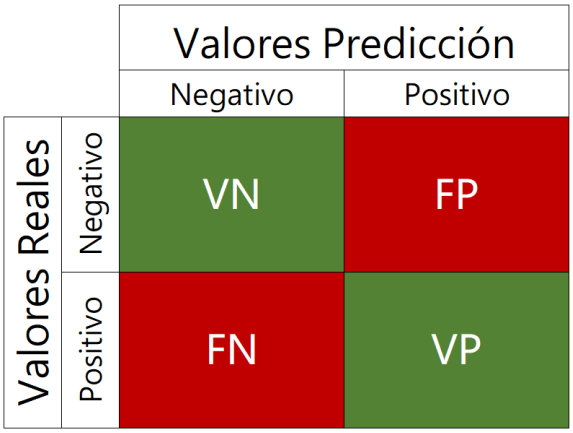


Figura 4.1: Matriz de Confusión.

Donde,

- VP: Verdadero Positivo (Valor predicción = 1; Valor real = 1)
- FN: Falso Negativo (Valor predicción = 0; Valor real = 1)
- FP: Falso Positivo (Valor predicción = 1; Valor real = 0)
- VN: Verdadero Negativo (Valor predicción = 0; Valor real = 0)

En base a la matriz de confusión se desprenden las métricas de rendimiento tal como sigue.

$$Acccuracy = \frac{VP + VN}{VP + FP + FN + VN}$$

$$Sensitivity = \frac{VP}{VP + FN}$$

$$Precision = \frac{VP}{VP + FP}$$

$$Specificity = \frac{VN}{FP + VN}$$

$$Balancedaccuracy = \frac{Sensitivity + Specificity}{2}$$

Con la finalidad de cumplir con el objetivo declarado al comienzo de esta sección, se requiere tener un buen manejo de las métricas recién definidas, puesto que cada una ataca un comportamiento en particular. Para el caso de estudio se necesita alcanzar una buena tasa de etiquetado de los clientes activos que tienen una alta probabilidad de fugarse (Valor predicción = 1), de tal forma que se pueda identificar con anticipación a los clientes que en el futuro cercano entrarán en la clasificación de un cliente inactivo, o sea tener un *recall* lo más alto posible.

Si bien, identificar a los clientes activos que potencialmente se fugarán es fundamental en un modelo de predicción de fuga, a nivel de negocio y como requerimiento del solicitante, se quiere lograr construir un proyecto que permita ser proactivos de cara a la fuga de clientes, pero que al momento de implementarlo se puedan minimizar los riesgos. Al hablar de riesgos de implementación se refiere principalmente a los costos elevados que puede incurrir la empresa en estudio producto de identificar erróneamente. Esto se relaciona directamente con el concepto de falso negativo, es decir, que el modelo identifique a un cliente activo como un cliente que con una alta probabilidad se fugará, sin embargo, este cliente realmente seguirá operando (cliente activo). Para poder mitigar este riesgo, se busca alcanzar valores elevados de *precision* y *specificity*. A pesar de que la métrica de *specificity* sea más precisa en lo buscado, se trabajará con la métrica de *precision* de modo que se pueda balancear el resultado objetivo.

A nivel de métricas de rendimiento, se mencionaron tanto a *recall* como a *precision*, que son los valores que se monitorearán con mayor énfasis para poder evaluar los modelos. Además de estas dos métricas se pondrá atención al *accuracy,* la cual muestra a nivel global con qué tasa el modelo evaluado predice comparado con la realidad. Por ende, se estará haciendo constante seguimiento de estas tres métricas para poder evaluar los modelos respecto a las métricas calculadas en base a la matriz de confusión.

Adicionalmente, otra forma de comparar el desempeño de los modelos es a través de la curva ROC [14]. La cual muestra, para distintos niveles de umbral de probabilidad de clasificación, la interacción entre True Positive Rate (TPR) o Recall y False Positive Rate (FPR) o [1-Specificity], generando una curva en dos dimensiones que va entre 0 y 1 en ambos ejes. Por lo tanto, como se quiere siempre tener un etiquetado lo más preciso posible, mientras la curva (ver Figura 4.2) esté más cerca del 1 o en su defecto la curva toma una forma lo más cercana a generar un cuadrado, mejor será el desempeño del modelo para clasificar correctamente. O bien que el área bajo la curva (AUC) sea lo más cercana a 1.

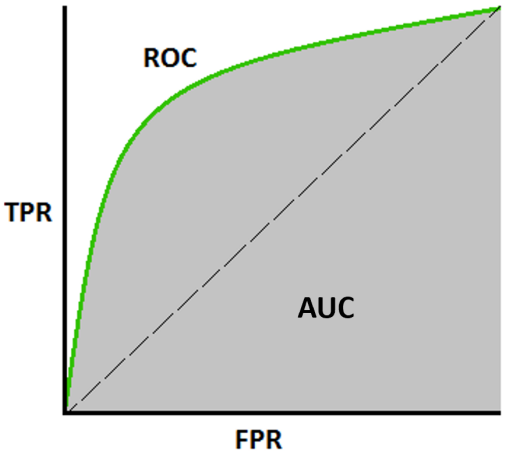


Figura 4.2: Gráfico de True Positive Rate (Recall) vs False Positive Rate (1-Specificity) - Curva ROC.

Según [14], en el caso que se cumpla que el valor del área bajo la curva sea igual a $1/m!$ (donde m es la cantidad de clases que presenta la variable dependiente del modelo, en este caso $m = 2$) la métrica AUC será igual a 0.5 o visto de otra manera, será igual que se etiquete al azar. Gráficamente, eso se vería representado de tal forma que la curva ROC sea igual a la diagonal punteada de la imagen.

En base a todas las herramientas de evaluación expuestas, se podrán comparar [15] de manera objetiva los modelos construidos, a fin de que se pueda cumplir el objetivo propuesto en esta sección.

Capítulo 5

Metodología

Para alcanzar los objetivos propuestos en este trabajo de título, se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [16]. Este tipo de metodología es una adaptación de un método genérico que se implementa en la ciencia de datos llamado Knowledge Discovery in Databases (KDD), donde se sigue una serie de etapas o pasos en un orden determinado con el objetivo de obtener información valiosa extraída desde una base de datos. Más aún, la finalidad del método CRISP-DM es incorporar la comprensión del negocio a la aplicación de modelos matemáticos. De esta forma, se describe a continuación cada una de las seis etapas de la aplicación de esta metodología, las cuales se pueden apreciar en la Figura 5.1. Sin embargo, la metodología involucra más etapas que las que se consideran en la metodología CRISP-DM, ya que el trabajo de título está compuesto por más etapas que deben desarrollarse. Estas estarán agregadas junto a lo mencionado anteriormente.

Figura 5.1: Diagrama de flujo metodología CRISP-DM.

5.1. Comprensión del negocio

Con la finalidad de comprender el funcionamiento del negocio se estudiarán en detalle las funciones de la Gerencia Comercial, de tal forma que se entienda con mayor profundidad las definiciones que la empresa tiene de un cliente fugado, qué condiciones impone para establecer esas definiciones, conocer los indicadores principales que ayudan a asimilar mejor las causas del problema planteado anteriormente; y las decisiones que se toman en base a los criterios mencionados. De esta manera, se podrá tener un manejo y construcción más guiado de las herramientas que se desarrollarán para solucionar el problema planteado, consiguiendo una interpretación más aguda, brindando mayor utilidad práctica al proyecto según los objetivos y lo esperado por el solicitante.

5.2. Comprensión de los datos

Al igual que en el punto anterior, esta etapa tiene por objetivo el entendimiento acabado, pero en esta ocasión sobre los datos e información que tiene almacenada la empresa en sus bases de datos. Con el fin distinguir claramente cuáles son los tipos de formatos en que los datos están almacenados (fecha, texto, numérico, etc), qué tipo de información contienen (características de clientes, datos transaccionales, etc) y, por último, interpretar correctamente esa información, relacionándola con la comprensión del negocio. De modo que el desarrollo de modelos matemáticos que contengan esa información como input tengan una relación directa y lógica con la operación de la empresa.

5.3. Preparación de los datos

Una vez realizado todo el estudio del negocio y su información almacenada por la empresa, se debe establecer una relación entre estos datos con lo que se quiere construir más adelante, o sea el modelo de predicción de fuga. Es decir, que se debe adaptar la base de datos a los modelos futuros, pues dependiendo del tipo de teoría o modelación que se esté implementando, distintos serán los requisitos que esa teoría exigirá para lograr desarrollarla de una manera adecuada. Concretamente, este apartado se refiere a hacer tratamiento de *missing values*, transformación de variables (normalización, estandarización, etc), selección de variables relevantes para los modelos, definición y acotamiento de la base de datos según la ventana de tiempo que se requiere analizar o alguna otra modificación a la base de datos original que el desarrollo del trabajo de título demande.

5.4. Modelamiento

Con el estudio del negocio y los datos, junto a su preparación realizados, se tiene el input necesario y en condiciones para poder iniciar la construcción de los modelos de predicción de fuga. Se habla de modelos en plural, puesto que existen dos opciones que se proponen para estudiar el problema planteado, un modelo de *regresión logística* y uno de *árbol de clasificación*. Este proceso, al igual que en los apartados anteriores, se realizará mediante lenguaje de programación Python y, en caso de requerirlo, lenguaje SQL, para manejar las bases de datos y extraer o modificar lo que sea necesario. En el desarrollo de los modelos se utilizarán librerías especializadas para construir modelos matemáticos de machine learning.

Este paso tendrá una modalidad iterativa, pues es importante ir probando los datos que se tienen de input, su funcionamiento con el modelo elegido, la coherencia de los resultados según lo que se sabe del negocio y la optimización de los hiperparámetros de los modelos para conseguir el mejor rendimiento posible según corresponda.

5.5. Evaluación

Dado el proceso de iteración en la construcción de los modelos anteriormente mencionados, se pasa a la fase de evaluación, es decir, calcular métricas de desempeño de los modelos de modo que los resultados sean consistentes con lo propuesto y que su desempeño sea lo más alto posible. Por tanto, esto será evaluado cuando esta etapa esté en pleno proceso, acompañado de la retroalimentación directa del solicitante. Si bien, este es el procedimiento estándar de evaluación de modelos, dado que se tienen propuestos dos modelos distintos, será muy relevante en este punto hacer una buena comparación de los modelos. Es ahí donde se presentan desafíos importantes como balancear los rendimientos versus complejidad, supuestos versus rendimiento, entre otras combinaciones que se descubran en el proceso. Para finalmente tomar la decisión de elegir un modelo por sobre otro para el proyecto.

5.6. Implementación de los modelos

La implementación de los modelos de predicción de fuga se hará en base a la evaluación anterior. El objetivo de esta etapa es poder disponibilizar una herramienta tecnológica y lo más automática posible que permita evaluar la propensión de un cliente activo a que se fugue, anticipando el proceso actual de retención de clientes que entra en acción una vez que un cliente activo supera los 60 días (cliente inactivo) sin operar.

5.7. Construcción de propuesta de experimentación y su evaluación

Tal como se ha expuesto a lo largo de este documento, se construirá una propuesta donde se especifiquen los tipos de incentivos comerciales que se les proporcionen a los clientes activos identificados como potenciales fugados. El propósito de esta etapa del proceso es proponer una forma de que los clientes etiquetados tengan incentivos para seguir operando y no entren a la clasificación de clientes inactivos. Luego de tener la propuesta se estudiará financieramente la implementación de las medidas comerciales sugeridas, de modo que se tengan distintos escenarios con diferentes niveles de éxito. Adicionalmente, con la propuesta ya estudiada se diseñará un experimento para que en el momento de la implementación se pueda medir el impacto del proyecto. La finalidad de este último punto mencionado es poder dejar planteados los parámetros y suposiciones clave para poder llevar a cabo un experimento que tenga validez estadística y se puedan desprender conclusiones con una buena base.

5.8. Elaboración de plan de implementación

Luego de desarrollar los apartados anteriores, solo queda ordenar las ideas y documentar lo realizado. Para esto, se elaborará un plan de implementación del trabajo de título, puesto

que después de terminar el proceso de desarrollo debe quedar una herramienta usable por la empresa y que sea independiente de la presencia del memorista. Es por esto que se propone la construcción de este documento, el cual va a contener definiciones de los conceptos básicos utilizados dentro del proyecto, cómo se utilizan los modelos, qué consideraciones se debe tener a la hora de usarlos y su relación con el negocio; cómo elaborar un plan de aplicabilidad de estrategias comerciales de fidelización según corresponda el caso y, finalmente, cómo llevar a cabo todo lo anterior de manera autónoma para que este instrumento sea sostenible en el tiempo.

Capítulo 6

Desarrollo Metodológico

Según lo planteado en la sección anterior, donde se proponen etapas a realizar durante el trabajo de memoria, se desarrollarán las etapas acorde a la metodología CRISP-DM aplicadas al caso de estudio.

6.1. Obtención y comprensión de los datos

El paso inicial para comenzar con el trabajo de título fue la inmersión dentro de la empresa en estudio y el rubro en que ésta opera. En particular, tal como se mencionó anteriormente, el desarrollo se sitúa dentro de la Gerencia Comercial de la empresa Chita Spa, siendo el Gerente Comercial la principal contraparte del trabajo. Acorde al objetivo y el problema expuesto, es importante entender primero qué factores están presentes dentro del proceso del factoring. Para esto se profundizó en la compresión de procesos como la adquisición de clientes, tanto de marketing digital como a través de canales directos y el embudo de clientes en su extensión; el manejo de los datos de este proceso, la aplicación de estrategias comerciales reactivas de clientes fugados, entre otras dimensiones.

Dado que el propósito metodológico es construir, mediante técnicas de machine learning, un modelo matemático que permita predecir qué clientes pasarán a la categoría de inactivos (fugados), lo prioritario fue entender las fuentes de datos y el significado de cada uno de los campos que estas contienen. Las bases de datos a las que se pudo acceder están cada una orientadas hacia un objetivo específico de negocio. Los datos están estructurados desde las siguientes fuentes:

- *Base transaccional*: base de datos que contiene toda la actividad operacional que la empresa realiza con sus clientes. En relación al servicio de factoring, esta tabla muestra en cada una de sus observaciones las facturas cedidas por los clientes y que fueron parte de algún proceso en la empresa. En ella se pueden encontrar datos de tipo fecha (para marcar hitos de la operación), información básica del cliente y el deudor (Rut, Razón Social, clasificaciones internas, etc), datos que identifican a las personas de la empresa que estuvieron involucradas en la operación y datos númericos mostrando el detalle de la operación (Monto, Tasa de interés, Anticipo, Mora, entre otras). Incluso están incluidas operaciones que no pudieron completarse y que por algún motivo tales facturas fueron recedidas.

- *Base transaccional orientada a marketing digital*: base de datos que tiene una estruc-

tura muy similar a la descrita previamente. De hecho, tiene variables idénticas, con la diferencia de que ésta solo contiene información extra sobre los canales digitales de donde fueron adquiridos los clientes, cuando éstos se crearon como lead, ingresos, costos y margen que se obtuvo por cada factura operada, por ejemplo. Para tener una mejor representación de esto, las variables que ésta base de datos comparte con la *Base Transaccional* representan un 28.2 % de ésta última.

- *Base Demográfica – SII*: base de datos que no pertenece a la empresa en estudio, sin embargo, muestra información pública de carácter demográfico acerca de empresas que han iniciado sus actividades hasta el año 2020 (última actualización disponible en la página del Servicio de Impuestos Internos, desde ahora SII). Donde se pueden encontrar variables como la Razón Social, Tramo según ventas, Número de trabajadores, Región, Comuna, etc.

Si bien la tabla principal, la *Base transaccional*, contiene datos desde el inicio de las operaciones de la empresa estudiada, esto no ocurre con las otras dos tablas. Ejemplo concreto de esta situación es que la Base transaccional orientada a marketing solo contiene inconsistencias en los datos para años anteriores al 2020. Asimismo, sucede con la *Base demográfica – SII*, la cual tiene información histórica de empresas desde muchos años atrás, sin embargo, tiene campos muy necesarios para el desarrollo de la memoria sin información. Esto se da en casos que tales empresas dejaron de existir o empresas muy nuevas que aún no tienen suficiente data como para completar su información, como por ejemplo el nivel de ventas que es insumo para catalogar a una empresa según tamaño.

6.2. Análisis exploratorio

Tal como se mencionó en el punto anterior, se tiene tres fuentes principales de datos de donde se obtendrá la información necesaria para desarrollar el modelo de predicción de fuga. Para lograr centralizar la información en una única tabla se realizó un proceso de cruce entre las tres bases de datos, coincidiendo según el rut del cliente. Además, desde la base del SII no se extrajo completamente la información ya que existían columnas que no aportan información relevante para el caso de estudio. De esta manera, se extrajo solamente información relacionada al tamaño de las empresas según sus ventas, número de trabajadores, rubro económico, actividad económica, región y comuna. De esta forma, la base de datos logra tener una mayor caracterización o segmentación de los clientes y no solo por su comportamiento transaccional, lo que puede ayudar a hacer un análisis más robusto.

Una vez que el proceso de unión de bases de datos se llevó a cabo, como resultado se obtuvo una tabla con **63. 603 filas** y **88 columna**s. La temporalidad de los datos sigue la lógica de lo expuesto en la justificación problema, donde se extrajo desde el año 2018 hasta el año 2021. Así, se tiene una base de datos que contiene operaciones realizadas específicamente desde el **01/01/2018** hasta el **31/10/2021**, obteniendo datos de casi 4 años de operación. A pesar de tener datos hasta el mes de octubre de 2021, la base de datos que se utilizará para entrenar y testear el modelo solo contendrá información hasta el mes de septiembre de 2021. La razón de esa reducción se responderá más adelante cuando se desarrolle el tratamiento de la base que servirá de input de los modelos.

Dentro de las 63 mil y fracción observaciones que contiene el dataset, se tienen **2.874 clientes únicos** que operaron con la empresa Chita durante el período analizado, los que servirán para el estudio de la fuga. Desde otra perspectiva, se tienen **46 camadas**[4] de clientes distintas lo que puede ayudar a tener un punto de vista diferente de análisis. Sin embargo, el foco principal será el comportamiento por cliente y su evolución en el tiempo.

6.2.1. Revisión de distribución de variables

Luego de tener una visión general de cómo está compuesta la base de datos con la que se trabajará, se revisó la distribución de algunas variables demográficas para entender su distribución en torno a los clientes y entender de una mejor forma el escenario en que se realizará el caso de estudio. Para esto, primero se analizó la distribución de clientes, a nivel porcentual y nominal, según el tamaño de la empresa (segmento SII), como se puede ver en la Figura 6.1 y Anexo B.1, respectivamente.

Como se puede ver en la imagen, hay una tendencia mayoritaria de operación con clientes de segmentos **Micro** y **Pequeña**, concentrando el **90.7 %** clientes. Si bien, todos los clientes deberían estar identificados en una categoría según el SII, eso no ocurre como se puede ver en el segmento "Sin info". Este último representa a clientes que operaron con Chita, pero que luego dejaron de operar como empresas y el SII los dejó como inactivos y, por ende, borró información dentro de la actualización de la base de datos. El otro caso que existe, es que debido a que se tiene una actualización no tan cercana a la fecha actual (2020), hay algunas empresas que han iniciado actividades y estas no se ven reflejadas. Este problema de falta de información es tratado más adelante, donde se puede verla nueva distribución de clientes que se utilizará en el modelamiento.

[4] Una camada es un conjunto de clientes que se agrupan según su creación en la empresa en un determinado período de tiempo. Para el caso de Chita, las camadas a las que se hace mención son camadas mensuales.

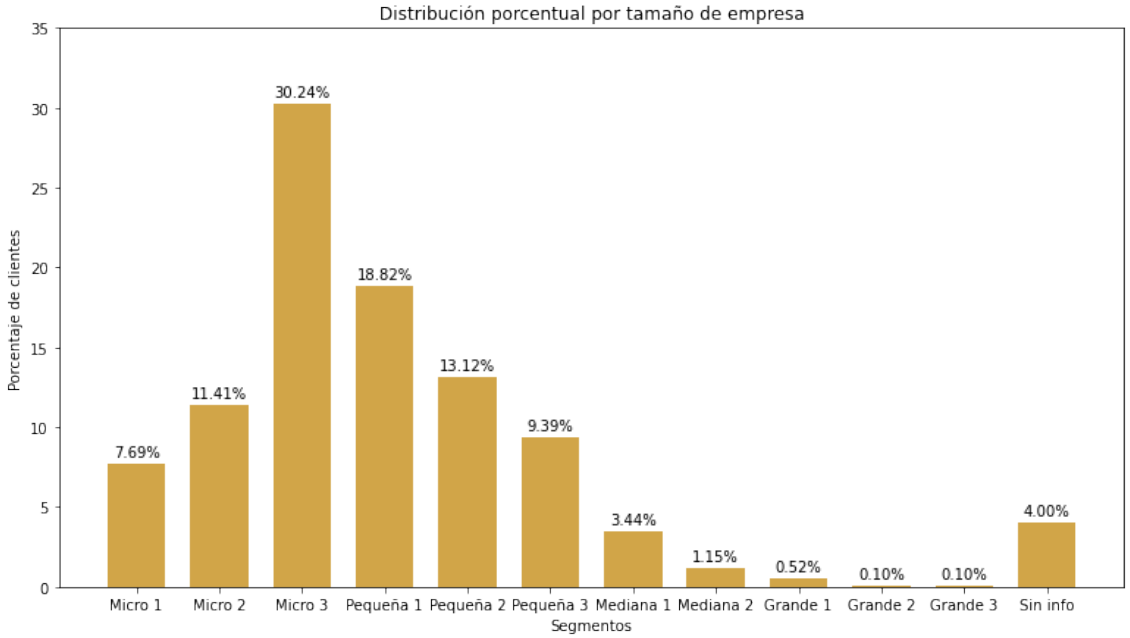


Figura 6.1: Gráfico de barras que muestra la distribución porcentual de clientes según tamaño de empresa (segmento según ventas, fuente SII).

Otra distribución importante a nivel demográfico de la base de datos es según el rubro al que pertenecen las empresas (ver Figura 6.2). Desde esta perspectiva se puede apreciar que un **95.5 %** de los clientes pertenecen a los diez rubros con mayor participación. Más aún, los tres rubros con mayor participación de clientes representan un **53.3 %**, es decir, aproximadamente **1 de cada 2** clientes pertenecen a uno de esos tres rubros. No obstante, cada uno de los rubros contienen dentro muchas formas de trabajo y una de las formas de indagar más allá es revisar la actividad económica en la cual están categorizados según el SII (Ver Anexo B.2).

Lo que se puede ver en la imagen es solo un resumen de las diez actividades económicas que más participación en clientes tienen, equivalente a un **41.8 %** del total de clientes. Si bien, se esperaba una mayor coincidencia entre los rubros y las actividades económicas, muestra que hay una gran diversidad de empresas con las que se trabaja.

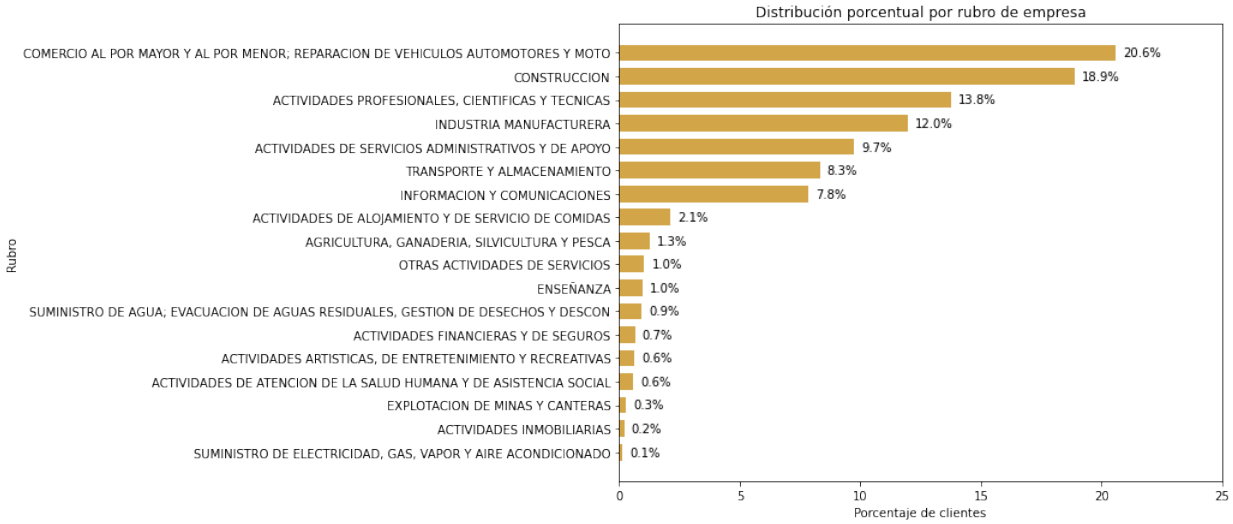


Figura 6.2: Gráfico de barras que muestra la distribución porcentual de clientes según rubro al que pertenecen las empresas (Rubro económico, fuente SII).

Finalmente, para terminar de entender cómo se caracterizan los clientes según su tamaño y giro al que pertenecen, es interesante poder entender cómo se distribuyen a lo largo del país, pues son solo clientes o empresas presentes en el mercado chileno. Este análisis se puede hacer a dos niveles con la información que se tiene, a nivel regional y comunal (ver Figura 6.3 y Anexo B.3).

Por otra parte, según las distribuciones por zonas demográficas, primero que todo, se puede ver una gran concentración en la región metropolitana (**55.7 %**) y el resto se distribuye se una manera bastante equilibrada a lo largo de las regiones. Dicho de otra manera, aproximadamente **1 de cada 2** clientes (empresas) pertenece a la región Metropolitana. Es por esta misma razón que al obtener el Top 10 de comunas con mayor participación en clientes, se tiene que **6 de las 10** son de esta la región Metropolitana. De hecho, las diez comunas con mayor participación representan a un **43.5 %** del total de clientes y en particular, las mejores 6 comunas de la región principal (RM) tienen una participación equivalente a un **31.9 %**, o sea casi un tercio de los clientes.

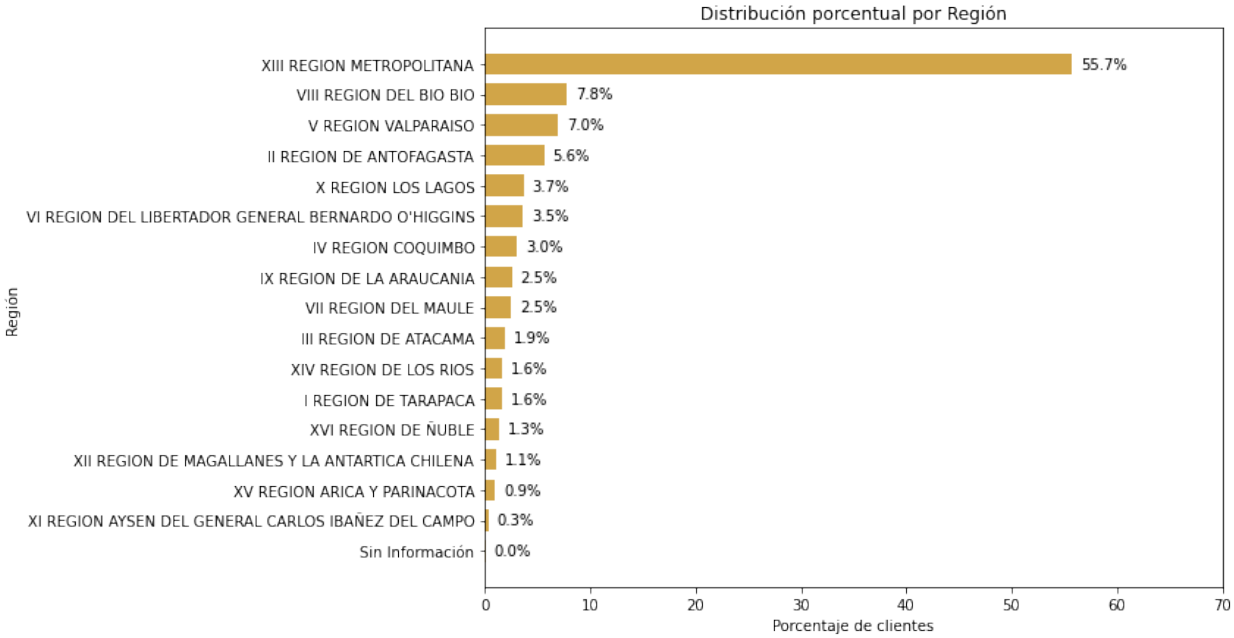


Figura 6.3: Gráfico de barras que muestra la distribución porcentual de clientes según región.

Otra perspectiva que se desprende del comportamiento transaccional de los clientes es cómo ha variado el margen capturado desde las operaciones de facturas respecto desde una mirada mensual. Esta variación está directamente relacionada con los montos y la cantidad de clientes que operaron con la empresa en el mismo mes, tal como se puede ver en la Figura 6.4. En la imagen se puede apreciar el valor porcentual que representa el margen obtenido y el monto de facturas total operado en cada mes, junto a la cantidad de clientes únicos operados en los mismos intervalos de tiempo.

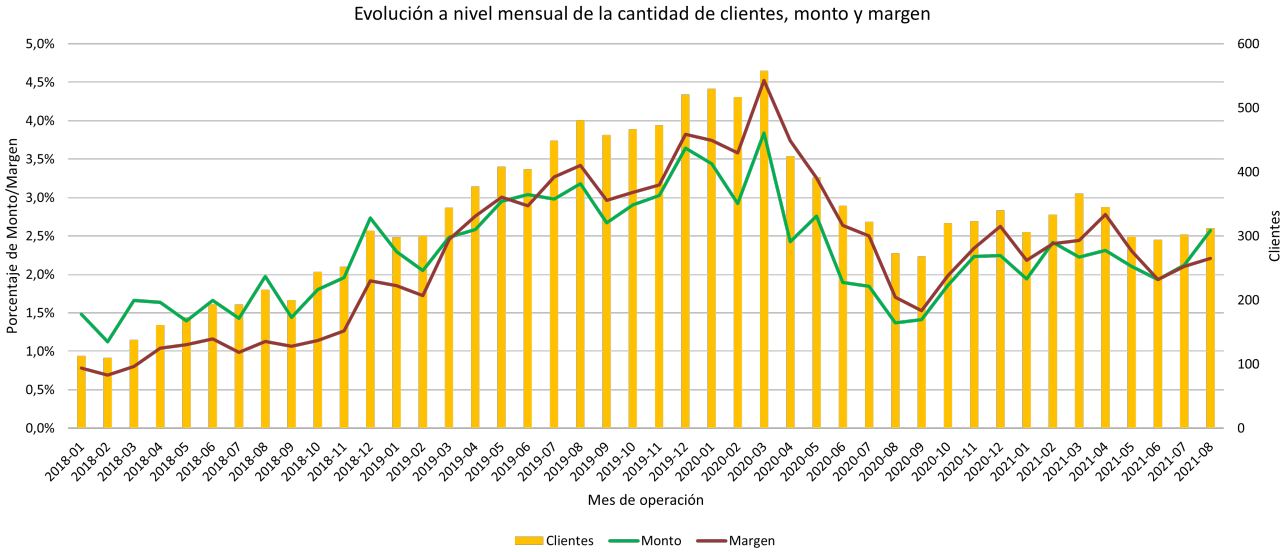


Figura 6.4: Gráfico combinado que muestra la distribución porcentual del margen obtenido mensualmente y el monto de facturas mensual total operado. Adicionalmente, muestra la cantidad de clientes únicos operados por mes.

En la imagen se puede ver cómo el nivel de margen obtenido mensualmente iba en alza, obteniendo su **peak en marzo de 2020**. Lo que se condice con la constante alza de la cantidad de clientes que, por consiguiente, deberían tender a operar montos de facturas (en total) más altos. Sin embargo, desde ese punto temporal se tiene una baja importante, llegando en solo seis meses al margen mensual mínimo obtenido (septiembre de 2020). Mes desde el cual se ha tenido un alza la cual llegó a una especie de estabilidad, pero no al nivel que se tenía en 2019 o principios de 2020.

6.2.2. Tratamiento de missing values y outliers

Hasta el momento se ha expuesto de manera descriptiva las distribuciones de las variables demográficas de los clientes presentes en el dataset construido. No obstante, existen algunos errores o algunas situaciones que hay que resolver para construir el modelo de predicción de fuga. Estos errores o valores anómalos que tiene la base de datos pueden ser del tipo outliers o missing values, los cuales se tratan a continuación.

6.2.2.1. Outliers

Para llevar a cabo el análisis de outliers, se toman en cuenta columnas relevantes[5] de la base de datos que en principio sean numéricas y puedan estar sesgadas por algunas observaciones con valores fuera de sus distribuciones respectivas. En particular, se analizaron cuatro variables, las que se muestran a continuación:

- Monto: monto de la factura
- Margen: margen obtenido de la operación de esa factura

[5] Columnas relevantes se le denomina, en este caso, a variables demográficas de los clientes o variables que demuestren comportamiento de operación o experiencia del servicio.

- Mora: monto de dinero que adeuda el cliente por operación de factura impaga
- Días mora: días que el cliente lleva sin pagar lo adeudado de esa factura

Con el fin de poder analizar sus distribuciones, primero que todo se verá a nivel de estadísticos si estas tienen un comportamiento un anómalo.

Tabla 6.1: Tabla que muestra las distribuciones de las variables Monto, Margen, Mora y Días mora. Fuente: Elaboración propia.

Estadístico	**Monto**	**Margen**	**Mora**	**Días mora**
Promedio	1.638.760	34.907	12.608	11
Desviación Est.	3.836.874	78.737	108.804	40
Mínimo	1.200	-53.098	0	0
Percentil 25	246.962	8.679	0	0
Mediana	620.599	20.743	0	0
Percentil 75	1.633.975	39.567	5.473	11
Máximo	227.773.800	12.997.520	12.837.530	1.216

De lo mostrado en la Tabla 6.1, se puede ver que hay algunas distribuciones que tienen valores mínimos menores a cero, lo cual no hace sentido para esa variable o valores máximos muy separados de la distribución (ver Anexos B.4-B.7). Es por ese tipo de situaciones que se decidió eliminar filas de la base de datos que no cumpliera con las definiciones de las variables o con los ejemplos expuestos.

Una vez eliminados los outliers de las variables expuestas, la base de datos queda con un total de **62.301 filas** y **88 columnas**.

6.2.2.2. Missing Values

Luego de resolver el problema de outliers es importante revisar desde otro punto de vista posibles errores de la base de datos. Para esto, primero se revisó si existen columnas relevantes con falta de información o missing values.

En este caso, la base presenta tres columnas relevantes que contienen missing values. Estas columnas o variables son:

- Región
- Comuna
- Tramo según ventas

Dado que estas variables no son numéricas, sino que son categóricas, resulta incompatible aplicar una técnica de imputación de datos simple donde se complete con el promedio, mediana, etc. Para este caso y dada la naturaleza de la columna, una de las opciones que puede representar de mejor manera los valores faltantes es mediante técnicas más complejas como una regresión lineal, que es la elegida para realizar este trabajo.

Tabla 6.2: Tabla que muestra cantidad nominal y porcentual de missing values para las variables Región, Comuna y Tramo según ventas.

Variable	Cant. Missing Values	Porc. Missing Values
Región	2	0,003 %
Comuna	2	0,003 %
Tramo según ventas	1.061	1,751 %

Para este fin, se calculó la cantidad porcentual y nominal de las observaciones con missing values para cada una de las tres variables.

En base a los resultados que entrega la Tabla 6.2, se puede establecer que las observaciones con falta de información para las columnas Región y Comuna son prescindibles en términos del peso que suman versus la base de datos completa. De modo que esas dos observaciones con missing values fueron eliminadas definitivamente. Por otro lado, los missing values de la variable Tramo según ventas, si bien en peso versus el dataset completo no es un gran porcentaje, son un poco más de mil observaciones que pueden imputarse y agregar valor al posterior modelo de predicción de fuga. Por lo tanto, se decidió ejecutar la técnica de regresión lineal para imputar los datos de esta columna, obteniendo la distribución mostrada en la Figura 6.5. Si bien, esta nueva distribución no muestra mayores cambios, ya no se tiene a ningún cliente sin la información del tamaño de la empresa.

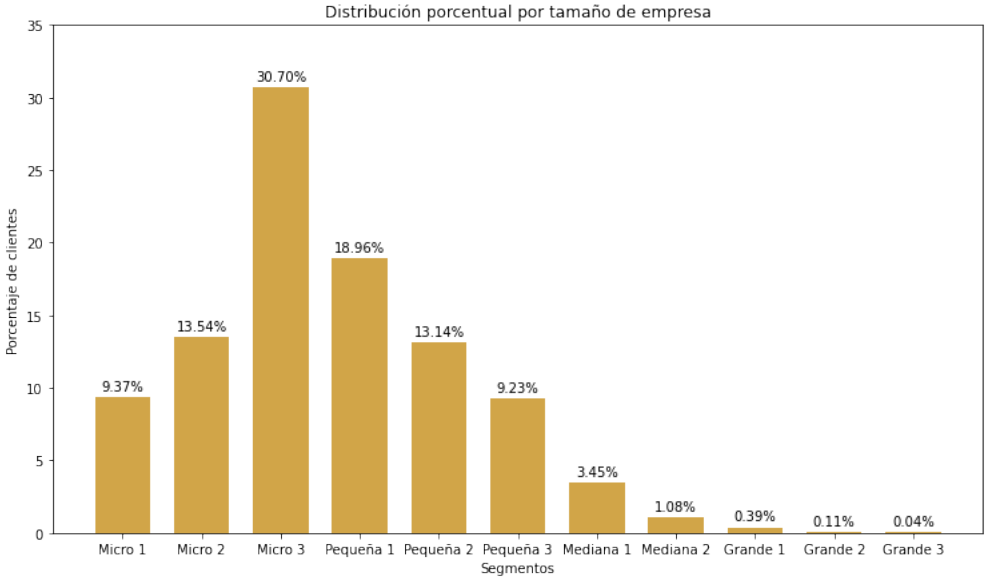


Figura 6.5: Gráfico de barras que muestra la distribución de los clientes por tamaño de empresa según el SII.

6.2.3. Reestructuración de la base de datos

A pesar de tener una base de datos con aproximadamente 63 mil observaciones y 88 variables, dentro de la misma no existe ninguna métrica que determine si un cliente está fugado o no. De hecho, por construcción de la base de datos, no se puede pasar a un modelo de predicción de fuga en base a la estructura de la base de datos original, puesto que la granularidad de este es a nivel de facturas y lo que se busca es tener una fuente de información que almacene

el comportamiento a nivel de cliente en un espacio temporal determinado (datos de panel). Es por esta razón que la base de datos ya existente sirvió como input para la construcción de la base de datos final, la cual es la definida para ser aplicada al modelo de predicción de fuga.

La base de datos fue construida bajo la lógica de hacer un seguimiento mensual al comportamiento de los clientes, o sea construir un dataset con una estructura de datos de panel. En este sentido, como se dijo al comienzo del análisis descriptivo, se tienen 46 meses de operación, de los que se verán reflejados solo 44 en el nuevo dataset. Esta reducción se realizó debido a la lógica que se quiere aplicar para entrenar los modelos de clasificación, donde el seguimiento de la etiqueta de fuga va a estar ligada a los datos de un mes hacia el futuro. Por ejemplo, las etiquetas de fuga que estarán presentes en el mes de enero del año 2021 serán las que se derivan del comportamiento de los clientes en el mes de febrero de 2021. Esta lógica se basa en poder tener para cada rut el comportamiento futuro de los clientes y que el modelo aprenda de las métricas que tiene en el mes original encontrando patrones que den indicios de una potencial fuga. Al mencionar esta lógica la base de datos debería reducirse en un mes (quedando en 45 meses de datos), sin embargo, quedan 44 meses debido a que se guarda el mes 45 para usarlo como validación del modelo elegido para hacer la predicción de fuga.

El inicio del análisis empieza en enero de 2018, contabilizando en ese periodo todos los clientes que operaron por primera vez en ese mes, junto a los rut's que están dentro de la base de datos que tienen una fecha de primera operación previo al inicio de 2018 (primeras operaciones del año 2016 o 2017). Luego, a medida que los meses avanzan, se van agregando los clientes que operan por primera vez en dicho periodo. De manera que, desde una mirada macro, la cantidad de clientes va aumentado cada vez que se ingresa a un mes nuevo. Sin embargo, al mismo tiempo que se agregan clientes, algunos se van eliminando de la base de datos en base a su Recencia o días que han pasado desde su última operación. Cabe destacar que el seguimiento se realiza el último día de cada mes (ejemplo: 31/01/2018, 28/02/2018, ..., 31/08/2021) para las variables no transaccionales como la Recencia, Antigüedad u otras variables similares que se calculan a partir de esa fecha. Mientras que para las variables transaccionales como el Monto, Días de mora, Frecuencia, entre otras, su seguimiento es acumulado en el tiempo. Esto quiere decir que, si un cliente ha operado 10 veces en total con chita y solo 1 vez en el último mes, quedará registrado el valor acumulado (10) en el último mes y no la frecuencia operada en ese mes.

La decisión de tener una lógica como la recién expuesta, surge de repensar lógicas aplicadas anteriormente que finalmente no calzaban con el propósito de la historia que debían contar. Es por esto que se tomó la decisión de usar las etiquetas de fuga de un mes hacia el futuro y de acumular los valores de las variables transaccionales, de tal forma que pudiera diferenciarse de mejor manera entre clientes activos o fugados.

Ya mencionada la lógica de construcción de la nueva base de datos, es relevante mostrar las variables presentes en ella, pues la cantidad de columnas difiere bastante de la base original. Para este fin, se seleccionaron variables que pudieran explicar el comportamiento o experiencia de los clientes en sus operaciones de manera agregada e información demográfica relacionada a cada uno de ellos. Así, en un dataset intermedio se seleccionaron las siguientes 8 columnas presentes en la Tabla 6.3:

Tabla 6.3: Tabla que muestra cada variable resultante del procesamiento inicial de las bases de datos. Fuente: Elaboración Propia

Variable	**Formato**
Rut Cliente	Numérica
Tramo según ventas	Numérica
Rubro Económico	Cadena
Actividad Económica	Cadena
Región	Cadena
Comuna	Cadena
Fecha Primera Operación Cliente	Fecha
Período Operación	Cadena

En base a este listado de columnas seleccionadas, se hizo un seguimiento mensual tal como se mencionó en la lógica de construcción (datos de panel), donde se calcularon métricas agregadas. Así como se dijo en secciones anteriores, se considerará como cliente fugado a un cliente que cumpla con que el valor de la Recencia esté entre 60 y 180 días. Dicho de otra forma, se toma la decisión de eliminar todos los registros marcados como perdidos, quedando una base de datos con **31.068** filas y **12** columnas.

	Rut Cliente	Tamaño	Rubro	Monto	Dias mora	Frecuencia	Recesiones	Hora giro	R/A	RM	Mora	Fuga
31067	[redacted]	Pequeña 3	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	230768675	7.078431	51	3	9.62199	0.003555	1	1	0

Figura 6.6: Imagen que muestra una observación de ejemplo de la base de datos resultante previo al seguimiento de la fuga.

Luego de pasar por un proceso de depuración de la base de datos, se revisó la correlación entre variables explicativas (ver Figura 6.7). Para esto se toman las principales columnas que determinan el comportamiento de los clientes y se ejecuta la matriz de correlación, a la cual se hizo un filtro de sus valores para no sobresaturar la visualización. De esta manera, los cuadrantes que muestran color es porque tienen una correlación menor a -0,3 o mayor a 0,3.

Figura 6.7: Imagen que muestra la matriz de correlación de las variables que contiene la base de datos construida para aplicar el modelo.

Según lo que se puede ver en la matriz de correlación, las relaciones más fuertes que existen se relevan son entre Frecuencia-Monto y Frecuencia-Recesiones. En ambos casos se tiene una correlación positiva. Para Frecuencia-Monto se tiene una correlación de 0.65 y para Frecuencia-Recesiones se tiene una correlación de 0.54. Aunque los valores de esas correlaciones son un tanto elevados, según su significado y lo aportado por la comprensión del negocio, estas variables aportan información muy relevante para describir el comportamiento y experiencia de los clientes, lo cual puede aportar en la relación con la fuga. Debido a esto, las columnas Frecuencia, Monto y Recesiones se mantienen en la base de datos.

Como información final luego de la construcción de la base de datos final, un 46,4 % de las observaciones están etiquetadas como clientes inactivos o fugados. Lo que demuestra que la base de datos está desbalanceada pues existen más filas que muestran a clientes activos que fugados. Pero a pesar de estar no estar balanceada, el dataset muestra un desbalance leve respecto a casos clásicos de bases de datos desbalanceadas del estilo 90 %-10 % o incluso 95 %-5 %. Lo cual puede presentar la ventaja de no recurrir a técnicas de resampling .

6.2.4. Análisis de fuga versus variables explicativas

Posterior a la construcción de una base de datos acorde con el modelamiento, es fundamental realizar un análisis donde se puedan relacionar las variables explicativas que se tienen dentro del dataset con la variable dependiente que se utilizará, la fuga.

Primero que todo, hay que tener en conocimiento que cuando se esté hablando de fuga se está hablando directamente de clientes catalogados como inactivos (entre 60 y 180 días sin operar). De este modo, para comenzar con el análisis es necesario tener en consideración cómo ha evolucionado la tasa de fuga mensual en el periodo de análisis del caso de estudio (ver Figura 6.8).

Tal como lo muestra el gráfico y lo mencionado en el apartado de justificación del problema, la tasa de fuga ha tenido una tendencia a aumentar a lo largo del tiempo. De hecho, el cálculo de la tasa de fuga anual coincide bastante con el peak que se alcanzó mensualmente (**julio de 2020**). Punto temporal del que la tasa de fuga empieza a descender alcanzando su punto más bajo en marzo de 2021. Sin embargo, percibió otra alza, alcanzando para el último mes

de datos una tasa de fuga mensual de un **37 %**, equivalente a **241 clientes** que a fines de ese mes pasaron al estado de **inactivos**.

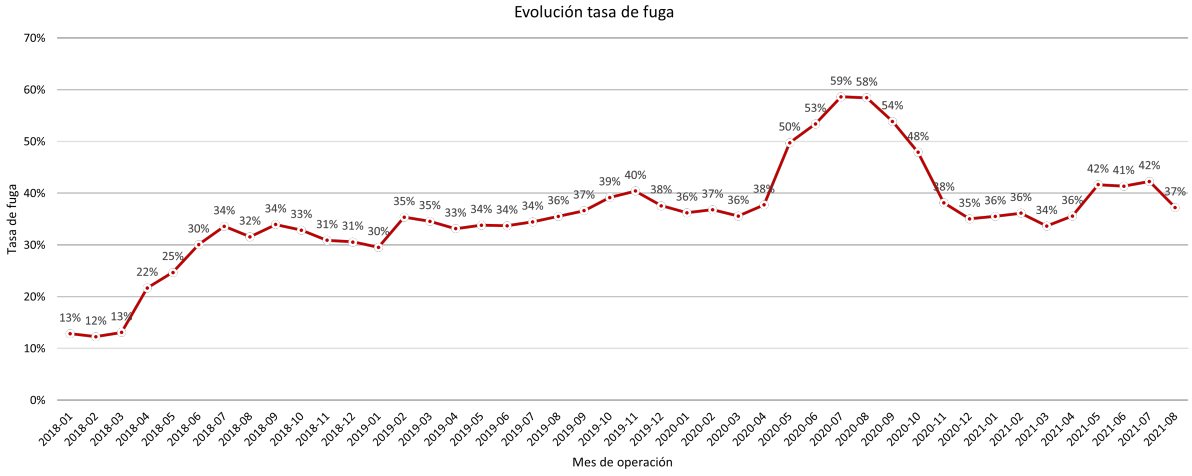


Figura 6.8: Gráfico de barras de muestra la evolución de la tasa de fuga por tamaño de empresa desde enero de 2018 hasta agosto de 2021.

También se estableció la relación entre el tamaño de las empresas con la fuga, pero en su evolución a través del tiempo. Para poder llevar esto a cabo, se calculó la tasa de fuga mensual para uno de los segmentos y luego se promedió a nivel anual quedando como resultado lo mostrado por la Figura 6.9.

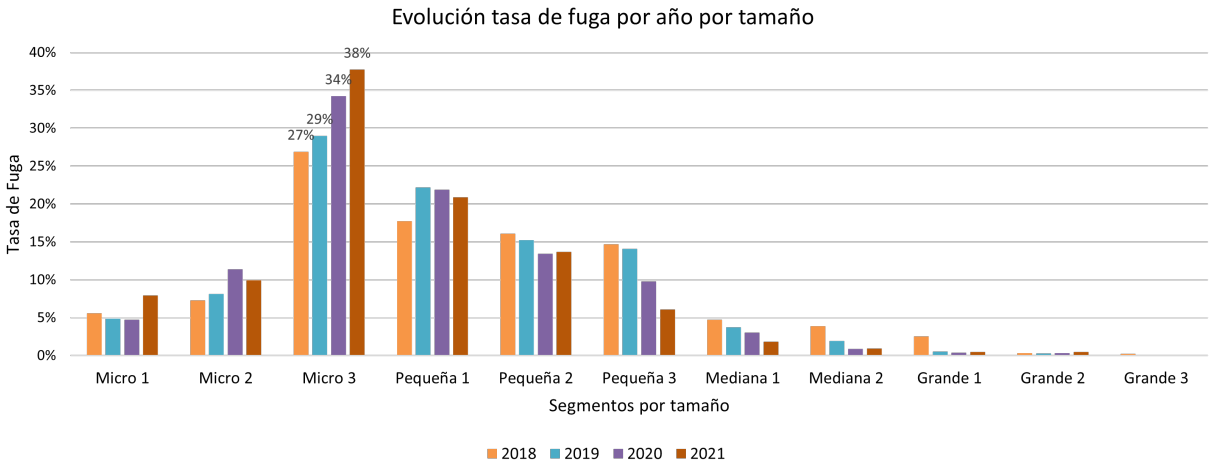


Figura 6.9: Gráfico de barras que muestra la evolución de la tasa de fuga por tamaño de empresa desde 2018 hasta agosto de 2021.

Según el gráfico de barras, se puede ver cómo algunos segmentos de tamaños han tenido un aumento de la tasa de fuga a lo largo de los años como las empresas **Micro 3** y otros que han disminuido como lo son las empresas **Pequeña 3**, **Mediana 1** o **Mediana 2**, las cuales muestran la evolución más marcada que el resto. En el caso de las clientes Micro 3, que son las empresas que tienen la mayor participación dentro de la empresa, se puede ver un aumento de **11 puntos porcentuales** desde el 2018 hasta agosto de 2021. Además, se puede ver a nivel global, que la fuga tiende a seguir una distribución similar a la distribución

de la cartera por tamaño. Esto último en algunos casos puede darse debido a una correlación entre la cantidad de clientes y la fuga y no solamente que por pertenecer a ese segmento se tendrá una mayor probabilidad de fuga (ver Anexo B.8).

Por otro lado, dado que aproximadamente la mitad de los clientes pertenecen a la región metropolitana (RM), en la base final se creó una variable binaria que marca con un 1 si la empresa tiene como residencia la RM y 0 si no. En esta misma línea, se estudió la relación entre pertenecer a la RM con la fuga de clientes, lo que se puede ver en la Figura 6.10. Es importante considerar que para esta visualización se consideraron solo a clientes que están marcados como fugados.

Tal como se ve en el gráfico, el impacto de pertenecer a la RM a lo largo del periodo de análisis no es constante y tiene variaciones bastante drásticas en algunos puntos. Por ejemplo, desde el mes de julio de 2018 (peak de mayor diferencia) se bajó hasta el mínimo de diferencia entre las tasas de fuga de ambos grupos la cual se dio en octubre de 2021. Dada la alta variación, un indicador que sirve para entender el comportamiento de esta variable sobre la fuga en el último periodo, es el promedio de la diferencia para los últimos ocho meses, el cual alcanza un 10 %. De este modo, para el año 2021 no se tiene una muestra clara de que ha influido en gran medida hacer la diferencia entre RM y fuera de ella con relación a la fuga.

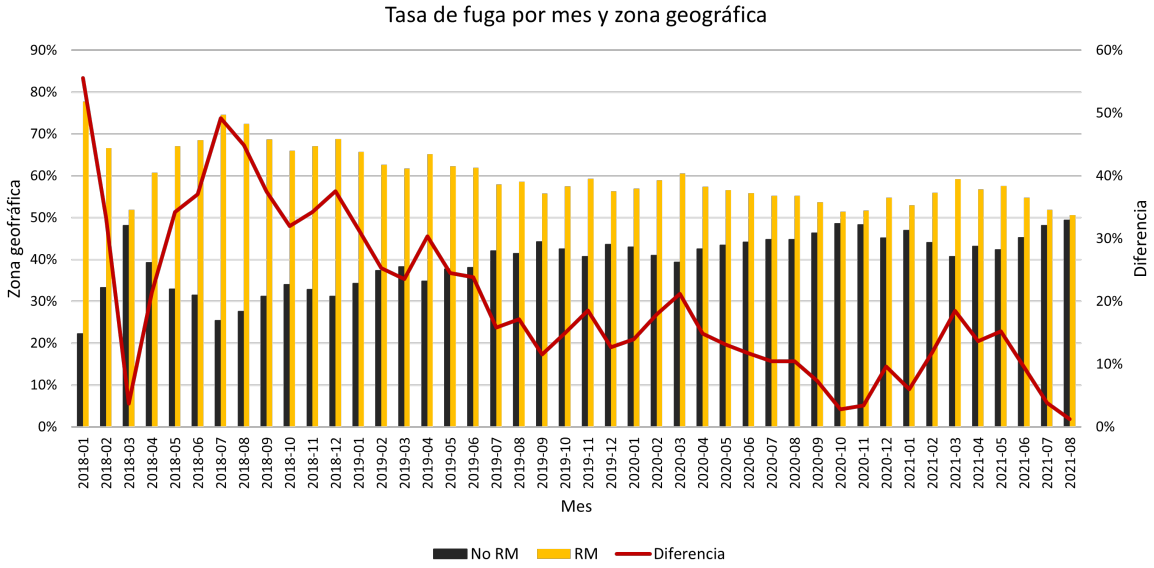


Figura 6.10: Gráfico de barras que muestra la evolución de la tasa de fuga por tamaño de empresa a través de los años de análisis.

Según las variables contenidas en la base final, aún quedan métricas importantes como la Recencia, la Antigüedad o la cantidad de clientes fugados por mes por revisar. La relación entre las variables mencionadas se establece gráficamente en la Figura 6.11. Como se mencionó previamente, la cantidad de clientes tendía a aumentar desde 2018. Sin embargo, luego del peak de principios de 2020 (**1.139 clientes**) esta métrica toma una curva descendente muy agresiva, repuntando solo a principios de 2021 (**742 clientes**).

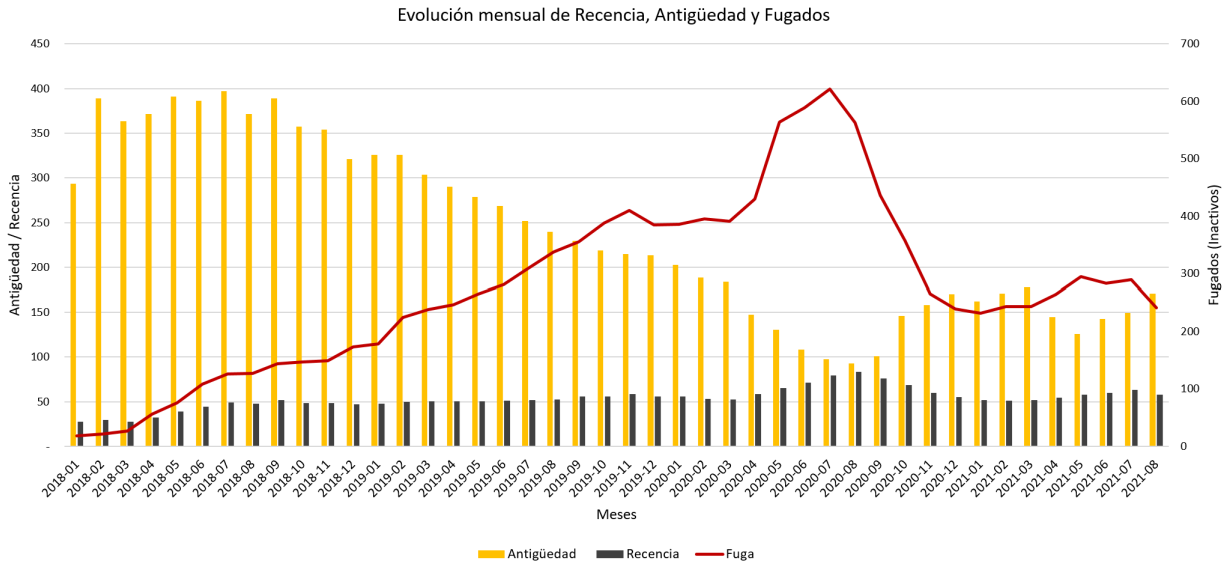


Figura 6.11: Gráfico de barras de muestra la evolución mensual de la Antigüedad, Recencia, clientes fugados (inactivos).

Basado en lo que se muestra en la imagen, a meses de empezar el contexto sanitario del COVID-19 la cantidad de clientes fugados toma una curva ascendente que comienza desde marzo de 2020 hasta llegar a su máximo en julio de 2020. Si bien no hay una causalidad clara de que la pandemia fue el único factor que causó esa tendencia en la curva de fuga, los números muestran que calzan temporalmente por lo que se puede establecer que fue un factor importante para la fuga de clientes. Por otra parte, se tiene la Antigüedad que, al contrario de las métricas anteriores, ha tenido una constante baja a lo largo del tiempo teniendo su punto más bajo exactamente el en punto más fuerte de la fuga (clientes inactivos). Dicho de otra forma, dado que había una alta tasa de fuga, la mayoría de los clientes activos durante ese periodo presentaban valores de Antigüedad baja debido a que como una gran parte de clientes se fugó, los clientes activos en ese momento (peak de fuga) operaron por primera vez en la empresa poco tiempo atrás. Junto a la Antigüedad está la Recencia que, en general, se ha mantenido constante durante todo el periodo de análisis, teniendo un movimiento promedio cercano a los **50 días**. Solo tuvo un alza significativa en el periodo en que los clientes inactivos tuvieron su máximo.

6.3. Modelamiento

Conforme al desarrollo anterior, la base de datos necesaria para poder aplicar los primeros modelos factibles de predicción de fuga está en condiciones de ser usada. Como se mencionó en la sección 4.2, se utilizarán dos modelos candidatos para predecir la fuga de los clientes de la empresa en estudio. Los modelos son una Regresión Logística y un Árbol de Clasificación.

Para ambos modelos de clasificación se utilizará el mismo set de variables inicial para probar y evaluar distintas configuraciones, las cuales se listan a continuación:

- Monto: monto acumulado de las facturas operadas hasta el mes de la observación.
- Días Mora: promedio acumulado de los días que el cliente lleva sin pagar lo adeudado

de las facturas operadas.

- Frecuencia: número acumulado de veces que un cliente opera hasta el mes de la observación.
- Recesiones: número acumulado de veces que un cliente presenta una factura redecida hasta el mes de la observación.
- Hora giro: promedio acumulado del tiempo que transcurre desde la cesión de una factura hasta el giro del anticipo medido en horas.
- R/A: ratio calculado entre la Recencia y la Antigüedad.
- RM: variable binaria que toma el valor 1 si pertenece a la Región Metropolitana y 0 si no.
- Mora: variable binaria que toma el valor 1 si el cliente presenta un monto mayor a cero respecto a la mora de las facturas y 0 si no.
- Rubro: variable categórica que determina el rubro del cliente. Al momento de incorporarse dentro de los modelos se divide en $n-1$ variables binarias dejando una de control (siendo n el total de categorías).
- Tamaño: variable categórica que determina el tamaño de la empresa (cliente). Al momento de incorporarse dentro de los modelos se divide en n-1 variables binarias dejando una de control (siendo n el total de categorías).

Previo a la ejecución de los modelos de clasificación, se divide la base de datos en dos grupos de similares características para poder validar de manera adecuada el desempeño de los modelos. Para realizar este proceso se toma el 80 % del dataset para utilizarlo como entrenamiento de los modelos y se deja el 20 % restante para su testeo. Con el fin de que el proceso cumpla con su propósito se debe verificar que la división de los datos cumpla con al menos la misma distribución de observaciones etiquetadas como activas y fugadas, para que no haya un desbalanceo mayor al inicial y los resultados se vean afectados. También, se verificaron las distribuciones de las variables categóricas de Tamaño y Rubro (ver Anexos 10-13). La distribución de la variable dependiente luego de la división de los datos se puede ver en Tabla 6.4.

Tabla 6.4: Tabla que muestra cantidad porcentual de observaciones que contiene cada subconjunto de datos para cada una de sus clasificaciones.

Variable	**Cant. Missing Values**	**Porc. Missing Values**
Entrenamiento	Activo	53.5 %
Entrenamiento	Fugado	46.5 %
Testeo	Activo	53.6 %
Testeo	Fugado	46.4 %

Según los datos mostrados en la Tabla 6.4 y en los Anexos B.9-B.12, se puede apreciar que las distribuciones de variables para la caracterización de las observaciones y de la variable dependiente conservan una distribución bastante similar entre los dos subconjuntos de datos.

Por lo tanto, es esperable que los resultados o el desempeño de los modelos de clasificación no se vean afectados por la división realizada.

Por otro lado, al aplicar los modelos, es fundamental realizar un proceso de selección de variables para encontrar la mejor combinación en base a las variables incluidas en el dataset que se usó como input. Para esto se recurrió al método de *Wrapping*, donde se probaron los métodos *Forward*, *Backward* y *Exhaustive*. Para la ejecución de ambos modelos se llevó a cabo este proceso de selección de variables, sin embargo, se incluyeron solo las variables numéricas (Monto, Días mora, Frecuencia, Recesiones, Hora giro, R/A, RM, Mora). En ambos casos, una vez realizado este proceso, se probaron dos combinaciones más de variables incluyendo variables categóricas (Tamaño y Rubro). De esta forma, se obtiene la mejor combinación de variables explicativas en relación a los valores de sus métricas de desempeño, tal como se podrá ver más adelante.

6.3.1. Regresión Logística

Lo primero que se debe revisar en este tipo de modelos de regresión son los supuestos que se requieren cumplir para que el modelo cumpla con su finalidad. Los supuestos principales a cumplir por este modelo son los siguientes.

Linealidad Esto se refiere a que los parámetros de la regresión que están dentro de la función de probabilidad no contengan formas funcionales que rompan la linealidad, es decir que se multipliquen o se apliquen formas funcionales más complejas como potencias, raíces, exponenciales, etc. En base a las restricciones de este supuesto, el modelo tal cual como se construyó cumple con la linealidad.

Multicolinealidad Este supuesto exige que no haya variables con altos valores de correlación, en particular, que no se incluyan variables que estén construidas en base a otra variable que esté dentro de la base de datos. Si bien se tienen las variables Frecuencia-Monto que tienen una correlación de 0.65, lo cual es un valor medianamente alto, se tomó la decisión de mantenerlas en el dataset ya que muestran un comportamiento que permite (juntas y por separado) caracterizar y diferenciar de mejor forma a los clientes. A pesar de tener un caso de correlación medianamente alta, se mitigaron los casos más importantes de dependencia en su construcción, por lo que el modelo cumpliría con este supuesto.

Dado que el modelo de regresión, según lo expuesto, cumple con los supuestos básicos que las regresiones exigen para tener un buen desempeño. Adicionalmente, es necesario saber cuáles son los hiperparámetros que se consideraron para la ejecución del modelo, los cuales se pueden ver a continuación.

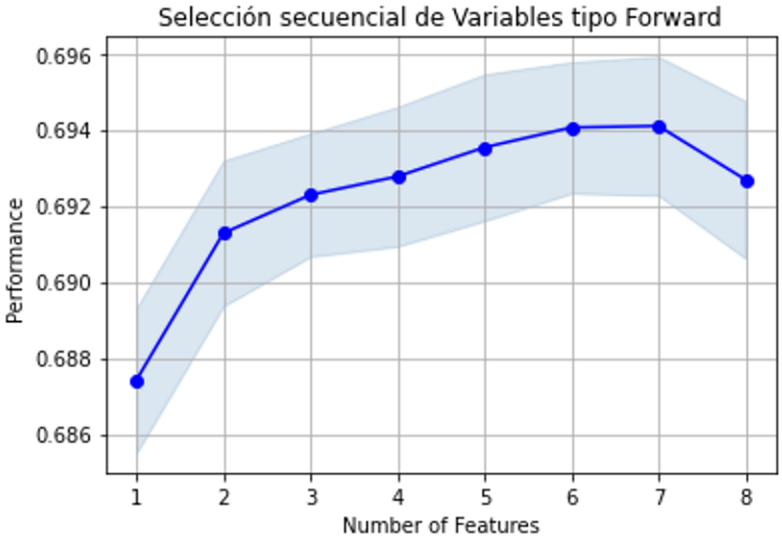
- Solver = liblinear
- Penalty = $l1$
- Max. Iteraciones = 10.000

Junto a esta configuración de hiperparámetros, se fija una semilla que permite obtener los mismos resultados cada vez que se ejecuta el modelo. La ventaja de tener una semilla fija es que se pueden probar distintas combinaciones de variables para evaluar qué combinación es

la mejor en términos de los resultados y métricas.

Tal como se mencionó anteriormente, se realizó la selección de variables en base al método de *Wrapping*. Para este modelo se utilizó específicamente el método de *Backward*, donde se inicia con la totalidad de las variables numéricas (en total 8), eliminando en cada una de las ocho iteraciones una variable que cumpla con tener el mayor valor de la métrica *accuracy* (ver Figuras 6.12). Como se puede ver en la imagen, la combinación que tuvo el accuracy más alto fue la iteración que contiene siete variables con un valor de **0.69**. Las variables que componen este subconjunto son las siguientes:

- Monto
- Días mora
- Frecuencia
- Recesiones
- Hora giro
- R/A
- Mora



	feature_idx	avg_score
8	(0, 1, 2, 3, 4, 5, 6, 7)	0.692685
7	(0, 1, 2, 3, 4, 5, 7)	0.694094
6	(1, 2, 3, 4, 5, 7)	0.694053
5	(1, 2, 4, 5, 7)	0.69353
4	(1, 2, 5, 7)	0.692766
3	(2, 5, 7)	0.692283
2	(5, 7)	0.691277
1	(5,)	0.687415

Figura 6.12: Gráfico y tabla que muestra los valores de acurracy para las distintas combinaciones de variables utilizando el método de *Backward*.

Con estas siete variables se crea la base del modelo de Regresión Logística con la que solo queda incluir y evaluar el rendimiento del modelo al incluir (juntas y por separado) las variables categóricas de la base de datos correspondientes al Tamaño y Rubro.

Una vez que se ejecutaron los distintos modelos, se tuvieron los resultados mostrados en la Tabla 6.5. Según lo que se puede ver en la tabla, el modelo con las métricas más altas es el que contiene la combinación de variables *Wrap + Tamaño + Rubro*. Este modelo tiene

Tabla 6.5: Tabla que muestra las métricas principales de todas las combinaciones de modelos probadas para la Regresión Logística, tomando el método de *Wrapping* como base.

Acurracy	**Precision**	**Recall**	**AUC**	**Variables**	**Modelo**
0.689	0.716	0.548	0.794	Wrap	RL
0.691	0.718	0.552	0.793	Wrap + Tam	RL
0.690	0.718	0.549	0.793	Wrap + Rub	RL
0.692	0.720	0.553	0.793	Wrap + Tam + Rub	RL

lo que se busca como objetivo de la modelación, tener un *Recall*, *Precision* y *AUC* altos. Que los valores de las métricas recién mencionadas sean altos indica que se está etiquetando correctamente tanto a clientes activos como inactivos.

En base a los coeficientes obtenidos de cada variable en el Anexo B.1. La variable usada que muestra el mayor coeficiente, y por tanto, la mayor influencia en la detección de una posible fuga es *R/A* ($\beta_{R/A} = 2.6$). Que ésta variable sea la más influyente en los resultados del modelo es esperable respecto a su construcción (*Recencia/Antigüedad*), donde la *Recencia* juega un rol fundamental. Luego está en segundo lugar el monto por conceptos de mora acumulado que tiene un cliente ($\beta_{Mora} = 6.0E-01$). El valor positivo del coeficiente de la *Mora* hace sentido de negocio con la fuga, puesto que un cliente que tenga un monto de mora tendrá una experiencia de servicio con etapas más inusuales y puede ser un factor relevante para fugarse. Por su parte, la *Frecuencia* acumulada tiene un valor negativo en su coeficiente ($\beta_{Frecuencia} = -5.5E-03$), lo cual deja en evidencia que mientras que un cliente menos veces acumuladas opere con la empresa de factoring mayor será su probabilidad de fuga. Al igual que con la Mora, la *Hora giro* ($\beta_{Hora\_giro} = -3.8E-03$) es una variable que demuestra concretamente la experiencia del cliente al contratar el servicio de la empresa de factoring, donde el valor negativo de su coeficiente representa que si el tiempo de entrega del anticipo es elevado, habrá una mayor probabilidad de fuga.

Por otra parte, se tienen a las variables dummies presentes en el modelo, donde primero para el *Tamaño* de las empresas se debe tener en cuenta que la categoría control o base es *Tamaño_Grande_1*. Esto es importante, pues los coeficientes son respecto al comportamiento de esta variable de control. De esta forma, los tamaños de empresa que tienen una mayor influencia son *Pequeña 3, Mediana 1, Mediana 2 y Grande 2*.

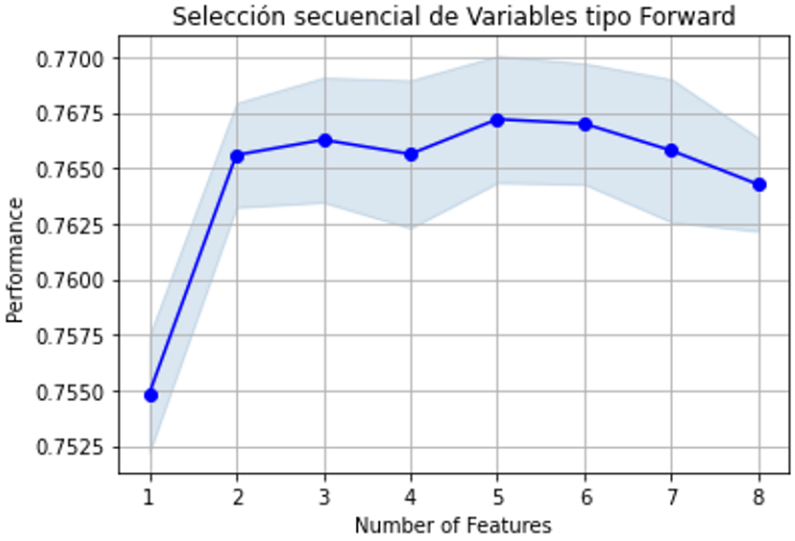
6.3.2. Árbol de Clasificación

Al igual que para el modelo de Regresión Logística, es necesario configurar los hiperparámetros del modelo y dejando una semilla fija para tener la opción de probar con distintas combinaciones de variables para que los cálculos se hagan de la misma forma, de modo que solo varíen las variables y sus resultados. Así, la configuración de hiperparámetros queda de la siguiente forma.

- Criterio = Gini
- Splitter = Best

- Max. Depth = 8

Análogamente al modelo anterior, se realiza el proceso de Wrapping, específicamente usando el método Forward. Este método funciona al revés que utilizado para la Regresión partiendo desde el modelo más simple (1 variable) hasta probar un modelo con el total de variables del dataset (ver Figura 6.13).



	feature_idx	avg_score
1	(5,)	0.754848
2	(0, 5)	0.765591
3	(0, 5, 7)	0.766275
4	(0, 1, 5, 7)	0.765631
5	(0, 1, 4, 5, 7)	0.767201
6	(0, 1, 4, 5, 6, 7)	0.766999
7	(0, 1, 3, 4, 5, 6, 7)	0.765792
8	(0, 1, 2, 3, 4, 5, 6, 7)	0.764263

Figura 6.13: Gráfico y tabla que muestra los valores de acurracy para las distintas combinaciones de variables utilizando el método de *Forward*.

Del método de Wrapping, donde se maximizó por la métrica accuracy con un valor de 0.77, se obtuvo como resultado la iteración 5 que contiene la siguiente combinación de variables:

- Monto
- Días mora
- Hora giro
- R/A
- Mora

Al ejecutar el método, los resultados se pueden visualizar en la Tabla 6.6.

Tabla 6.6: Tabla que muestra las métricas principales de todas las combinaciones de modelos probadas para la Regresión Logística, tomando el método de *Wrapping* como base.

Acurracy	Precision	Recall	AUC	Variables	Modelo
0.760	0.699	0.851	0.826	Wrap	AC
0.763	0.700	0.855	0.829	Wrap + Tam	AC
0.761	0.700	0.851	0.826	Wrap + Rub	AC
0.762	0.700	0.853	0.827	Wrap + Tam + Rub	AC

Según los resultados obtenidos del *Wrapping* mostrados en la Tabla 5, el modelo que tiene un mejor rendimiento en relación a sus métricas es el modelo *Wrap + Tamaño*.

6.3.3. Elección de modelo de clasificación

Posterior a la ejecución de ambos tipos de modelos con las combinaciones de variables descritas en los apartados anteriores, se obtiene un resumen de todos los resultados contenidos en la Tabla 6.6.

Tabla 6.7: Tabla que muestra las métricas principales de todas las combinaciones de modelos probadas, tomando el método de *Wrapping* como base.

Acurracy	**Precision**	**Recall**	**AUC**	**Variables**	**Modelo**
0.689	0.716	0.548	0.794	Wrap	RL
0.691	0.718	0.552	0.793	Wrap + Tam	RL
0.690	0.718	0.549	0.793	Wrap + Rub	RL
0.692	0.720	0.553	0.793	Wrap + Tam + Rub	RL
0.760	0.699	0.851	0.826	Wrap	AC
0.763	0.700	0.855	0.829	Wrap + Tam	AC
0.761	0.700	0.851	0.826	Wrap + Rub	AC
0.762	0.700	0.853	0.827	Wrap + Tam + Rub	AC

A raíz de éstos resultados, se establece que según las métricas que se presentan, el modelo elegido como el mejor para clasificar a clientes activos y fugados es el modelo de tipo *Árbol de Clasificación (AC)* en su variante *Wrapping + Tamaño*. Para ese modelo se tiene un rendimiento general representado por el *accuracy* con un valor de **76.3 %**. Luego, de las observaciones que en realidad son clasificadas como clientes fugados versus los que el modelo etiquetó como fugados se tuvo una tasa de un **85.5 %**. Como se ha dicho a lo largo de este documento, la tasa recién descrita (*Recall*) es muy relevante pues muestra que tan bien se identifican a clientes activos potenciales a fugarse. Además, está la métrica *Precision* que muestra, dentro de lo que el modelo predice como fugados, a los que realmente son fugados en una tasa de un **70 %**. Finalmente, según la tabla de resultados muestra un *Área Bajo la Curva (AUC)* de un **0.829** (ver Anexo B.13), que es el valor más alto dentro de los modelos probados. Esto indica que existe una buena relación en la identificación tanto de clientes activos como los fugados.

Una forma más visual de revisar estos resultados es mirando la matriz de confusión del modelo elegido. Como lo muestra la Figura 6.14 y como se mencionó en base a la tabla de resultados, el modelo elegido identifica con una tasa de un 85.5 % a los clientes fugados. Por su parte, identifica a los clientes activos con un **68.2 %**. Si bien el modelo etiqueta con una mejor precisión a los clientes fugados que a los activos, se tiene en general un buen rendimiento del modelo en base al resto de modelos probados.

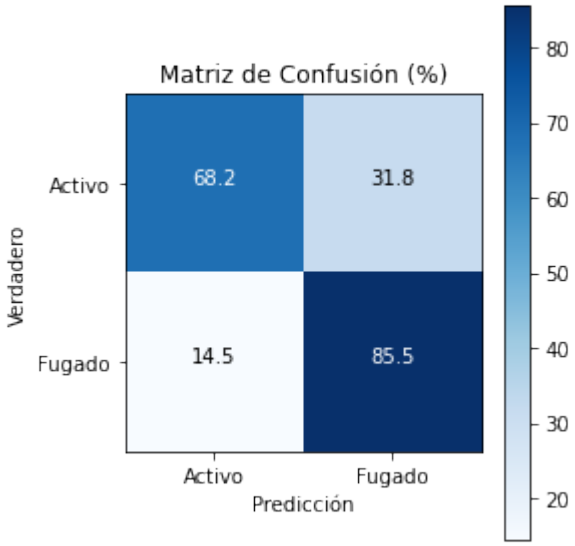


Figura 6.14: Matriz de Confusión del modelo elegido.

Dado que el modelo elegido es del tipo Árbol de Clasificación, se pueden obtener los valores de la importancia que tiene cada variable en la ejecución del modelo al momento de clasificar. La importancia recae en la capacidad explicativa de las variables sobre la varianza de las observaciones, es decir, las variables con mayor importancia tienen una mayor capacidad de discriminar entre los diferentes comportamientos de los clientes (activos o fugados). Esto se puede visualizar de forma más clara en la Figura 6.15.

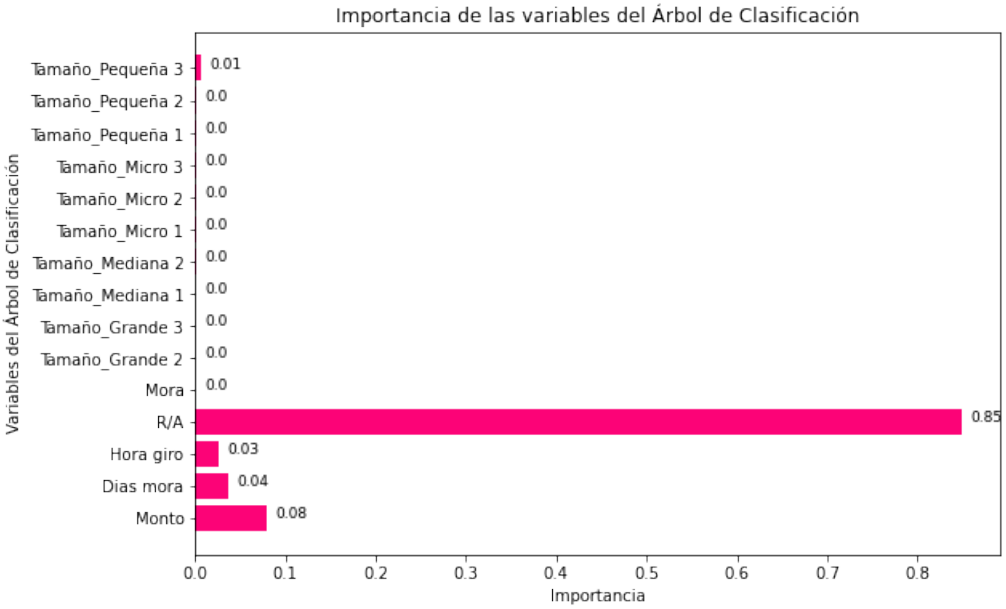
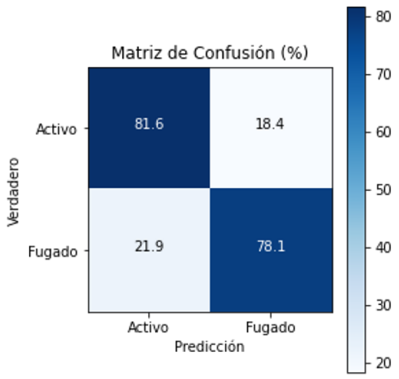


Figura 6.15: Gráfico de barra de muestra la importancia de las variables utilizadas en el modelo de clasificación elegido.

Basado en el gráfico de importancia, el top 3 de variables que más relevancia tienen a la hora que hacer la clasificación son *R/A, Monto, Días mora*. Si se analiza lo que representan detrás (y en general las variables incluidas en el modelo elegido), tienen mucha relación con su antigüedad, los días que tienen sin operar, el monto total de las facturas operadas y los

días de mora que en promedio tienen los clientes; que son variables que demuestran concretamente el comportamiento del cliente en las operaciones realizadas con la empresa en estudio. En particular, sobresale la importancia de la variable *R/A* que representa el ratio entre la *Recencia* y la *Antigüedad*, lo que es coherente con su significado puesto que la *Recencia* define directamente el estado de un cliente. En segundo lugar se tiene al *Monto* acumulado de las facturas operadas por un cliente, lo que da cuenta que dependiendo del tamaño del monto junto a la frecuencia de operación[6] pueden mostrar un comportamiento más recurrente o más espaciado. Las variables *Días de mora* y *Hora de giro* muestran concretamente el comportamiento y experiencia del cliente a la hora de hacer una transacción, por lo que es coherente que su importancia sea mayor que el tamaño de la empresa por ejemplo. La importancia de estas variables presentes en el modelo tienen mucho sentido de negocio en relación a la fuga de un cliente, por lo tanto, se interpreta como una confirmación de que el modelo está capturando comportamientos de negocio mostrados en los datos.

Por otra parte, en la sección 6.2 se mencionó que al inicio se contaba con 46 meses de operación, sin embargo, para los modelos se ocuparon solo 45. La finalidad de este proceso es poder, con un mes de datos que no se haya usado en el entrenamiento o testeo, validar el rendimiento del modelo elegido. Específicamente, se utilizó el mes de septiembre de 2021 para hacer esta validación, obteniendo los siguientes resultados (ver Figura 6.16).



Accuracy	Precision	Recall	AUC
0.801	0.752	0.781	0.877

Figura 6.16: Matriz de Confusión y tabla de resultados de la validación del modelo de clasificación elegido.

Según estos últimos resultados, se puede ver que el modelo elegido se comporta en un nivel bastante parecido (incluso mejor) al rendimiento que se pudo ver en las etapas anteriores. Este fenómeno muestra que el modelo es robusto en la clasificación, pues tiene una alta tasa de clasificación tanto de clientes activos como fugados. De hecho, para este caso la identificación de clientes activos fue mejor que la de clientes fugados, lo cual se alinea con la mitigación de riesgo para la implementación (menor inversión aplicada erróneamente en clientes mal etiquetados).

Con todo lo desarrollado a lo largo de este apartado, los resultados muestran que el modelo de clasificación elegido tiene sentido con el negocio en relación a las variables elegidas para ser parte de la predicción, la importancia de las variables elegidas y como en conjunto muestran

[6] El *Monto* y la *Frecuencia* son variables que se relacionan mucho en su interpretabilidad. Un *Monto* de factura acumulado alto se puede deber a una baja frecuencia de operación con montos de factura altos o una alta frecuencia con montos de factura bajos

cumplir con los objetivos propuestos para la etapa de modelamiento al clasificar e identificar a clientes activos y fugados.

6.4. Propuesta de medidas comerciales de retención

Según estos últimos resultados, se puede ver que el modelo elegido se comporta en un nivel bastante parecido (incluso mejor) al rendimiento que se pudo ver en las etapas anteriores. Este fenómeno muestra que el modelo es robusto en la clasificación, pues tiene una alta tasa de clasificación tanto de clientes activos como fugados. De hecho, para este caso la identificación de clientes activos fue mejor que la de clientes fugados, lo cual se alinea con la mitigación de riesgo para la implementación (menor inversión aplicada erróneamente en clientes mal etiquetados).

Con todo lo desarrollado a lo largo de este apartado, los resultados muestran que el modelo de clasificación elegido tiene sentido con el negocio en relación a las variables elegidas para ser parte de la predicción, la importancia de las variables elegidas y como en conjunto muestran cumplir con los objetivos propuestos para la etapa de modelamiento al clasificar e identificar a clientes activos y fugados.

- Tasa de Interés asociada a la factura
- Comisión asociada a la factura
- Porcentaje de Anticipo[7]

Estos instrumentos expuestos tienen directa relación con la segmentación que la empresa tiene del cliente que realiza la transacción, su clasificación de riesgo y el tamaño de la factura cotizada. De manera que se genera un precio dinámico dependiente de las variables recién mencionadas.

La propuesta de la oferta se hará en base a una combinación se descuentos realizados sobre los tres instrumentos expuestos. La oferta está orientada a que los clientes identificados y quienes son parte del proceso reoperen en un plazo corto de tiempo. Por esta razón se enviará la oferta vía correo y con apoyo del equipo de ejecutivos(as) a través de llamadas telefónicas, dando un plazo de 15 días para aceptar la oferta y operar con la empresa en estudio.

En particular, la oferta se detalla según la Tabla 6.8.

Tabla 6.8: Tabla que muestra el detalle de la oferta propuesta según los instrumentos expuestos.

Instrumento	**Oferta**
Tasa de interés	20 % dcto.
Comisión	15 % dcto.
Porcentaje de anticipo	Fijo en 98 %

[7] El Anticipo es el monto que se acuerda (en porcentaje respecto al monto de la factura) entre la empresa de factoring y el cliente para llevar a cabo el financiamiento.

Si bien más adelante en el trabajo de título se tratará la dimensión financiera de implementar esta oferta en conjunto con los resultados del modelo, faltan algunos antecedentes para poder hacer de manera adecuada los cálculos. Es por esta razón que se mostrará la diferencia del costo que pagaría un cliente con y sin la oferta, para tener una referencia del incentivo que se propone para los clientes.

Como referencia se tiene una factura de un monto de $1.000.000, realizado por una empresa clasificada como mediana y con una clasificación de riesgo cliente-deudor de A+/A+ (en el apartado 6.6 se profundizará en las clasificaciones mencionadas). De esta forma, en la Tabla 6.9 se muestran los valores de la comparación al aplicar la oferta propuesta, mostrando solo los costos de servicio.

Tabla 6.9: Tabla que muestra la comparación del costo de servicio.

Costo de servicio	**Tasa de interés (%)**	**Comisión($)**	**Total ($)**	**Diferencia ($)**	**Diferencia (%)**
Normal	1.53	24.990	74.970	-	-
Oferta	1.22	21.242	61.634	-13.337	-17.79

Como se puede ver en los cálculos mostrados en la tabla, en base a los costos de servicio se puede ver que un cliente (según las clasificaciones fijadas como referencia) podría percibir un descuento de un **17.8 %** en el costo de operar una factura con la empresa estudiada. Es importante mencionar que el porcentaje de descuento irá variando dependiendo de las clasificaciones del cliente y del tamaño de la factura que presente. Además, en esta comparación solo se consideró el costo de servicio, sin embargo, hay un beneficio adicional que tiene que ver con el instrumento no presente en la Tabla 6.9, el porcentaje de anticipo. Pese a que el promedio de porcentajes de anticipos en el año móvil (septiembre de 2020 hasta agosto de 2021) respecto a la ventana temporal que se tiene de los datos es de un 98 %, según los datos se ha financiado desde un 49 % hasta un 100 %. Por lo que fijarlo en un 98 % es una garantía de que este beneficio no será más bajo que el promedio.

De esta manera, se deja propuesta una oferta que tiene tres ejes de beneficios para los clientes en relación el costo que deben pagar por recibir el servicio de factoring y que da la opción de poder realizar la primera transacción hasta 15 días después del primer contacto con el cliente identificado como potencial fuga.

6.5. Diseño del experimento

Construir un diseño experimental es fundamental para medir empíricamente el efecto de identificar clientes y aplicarles medidas comerciales en relación al comportamiento de la tasa de fuga. Para llevar a cabo la construcción del diseño del experimento es necesario tener en cuenta el funcionamiento general respecto a los clientes.

En primer lugar, se define de manera global que se realizará una medición entre dos grupos de clientes. El grupo total de clientes que participarán en el experimento serán elegidos en base a su estado, es decir, se elegirán solo a clientes en el estado de activos. Se toma esta decisión debido a que los clientes activos son aquellos que tienen la opción de fugarse o seguir operando con la empresa, que es lo que se busca medir en base a los diferentes estímulos presentes en la experimentación.

En este proceso es necesario dividir a los clientes activos (pertenecientes a un periodo de tiempo determinado) en dos grupos, grupo *control* y *tratamiento (50 %-50 %)*. Dentro del grupo control no se aplicará ningún estímulo a los clientes activos seleccionados, o sea que permanecerán según la estrategia de la empresa en estudio, la cual es reactiva y se pone en marcha luego de que un cliente tiene el estado de inactivo (fugado para efectos del trabajo de título). Mientras que para el grupo de tratamiento se aplicará la oferta propuesta en la sección 6.4, de modo que en una ventana de tiempo a definir se haga seguimiento al comportamiento de esos clientes, para finalmente comparar la tasa de fuga entre ambos grupos.

La división de los clientes activos elegidos para el experimento entre el grupo control y tratamiento es un paso vital para el proceso y debe tener la atención necesaria. Una mala implementación en la selección de los miembros que irá a cada grupo implicaría en se incluyan sesgos dentro de la medición aumentando la probabilidad de error en las conclusiones de los test aplicados en la comparación. Es por esta razón que para hacer la división se buscará mantener la misma distribución de variables que caracterizan a los clientes como por ejemplo el Tamaño, Rubro o alguna otra variable que clasifique o describa a los clientes. Generar una correcta separación de los clientes activos busca que la probabilidad de que cada miembro (del total de clientes) pertenezca a algún grupo sea la misma.

Tal como se dijo previamente, el experimento medirá la diferencia de la tasa de fuga de ambos grupos. Para realizar esa medición se utilizará un test de proporciones, donde cada proporción de ambos grupos sea exactamente su tasa de fuga. Es así que se define la hipótesis nula (H_0) e hipótesis alternativa (H_1) como se puede ver a continuación.

$$H_0 : p_{control} \leq p_{tratamiento}$$
$$H_1 : p_{control} > p_{tratamiento}$$

Siendo $p_{control}$ la proporción de personas que se fugaron del grupo control y $p_{tratamiento}$ la proporción de persona que se fugaron del grupo tratamiento. Planteado de otra manera, la hipótesis nula (H_0) exige que la tasa de fuga de grupo control sea menor o igual a la tasa de fuga del grupo tratamiento. En cambio, la hipótesis alternativa (H_1) exige que la tasa de fuga del grupo control sea mayor a la tasa de fuga del grupo tratamiento. Dado que las hipótesis muestran desigualdades entre las proporciones de cada grupo, el test de proporciones será un test de una cola. Esto busca una mayor precisión en la medición del efecto, pues se quiere saber específicamente si el impacto de identificar a clientes activos potenciales a fugarse y darles un estímulo comercial repercute en que el grupo tratamiento tenga una tasa de fuga menor al grupo control.

Adicionalmente, uno de los puntos más importantes de construir el diseño del experimento que se quiere realizar es conocer el tamaño muestral que se requiere para conseguir un cierto nivel de confianza $(1 - \alpha)$ y poder estadístico $(1 - \beta)$. Para este experimento se fijará el nivel de confianza en un 95 %, mientras que para el poder estadístico se evaluará que valor fijar, el cual se elegirá en función del tamaño muestral requerido para distintos niveles de poder

(80-95 %) y de la cantidad de clientes activos que hay en la base de datos mensualmente.

El primer paso para fijar el valor del poder estadístico que se utilizará en este experimento es calcular el tamaño muestral. Para esto se utilizó un programa estadístico especializado en experimentos y test de hipótesis llamado *G*power* [17], elaborado y disponibilizado públicamente por la Universidad de Düsseldorf[8]. El cálculo del tamaño muestral depende del valor del nivel de significancia ($\alpha = 5\,\% = 0.05$), del poder estadístico ($1 - \beta$), del ratio de división que será igual a 1 pues el tamaño de los grupos del experimentos son el mismo y de las proporciones (tasa de fuga) de los grupos. En relación a las proporciones de los grupos, en el caso del grupo de control se tomó (para el último año móvil) la tasa de fuga mensual de los clientes activos que se fugan al siguiente mes, resultando una tasa de fuga promedio de **17.6 %**. Por su parte, la proporción del grupo tratamiento no se conoce, por lo que se fija un valor de tasa de fuga objetivo para comparar los grupos en un **10 %**. Es decir, el tamaño del efecto que se espera medir en el experimento es de **7.6 %**. Con estos datos configurados en el programa estadístico, se calculan los tamaños muestrales correspondientes a los distintos niveles del poder estadístico (ver Figura 6.17 y Anexo C.1).

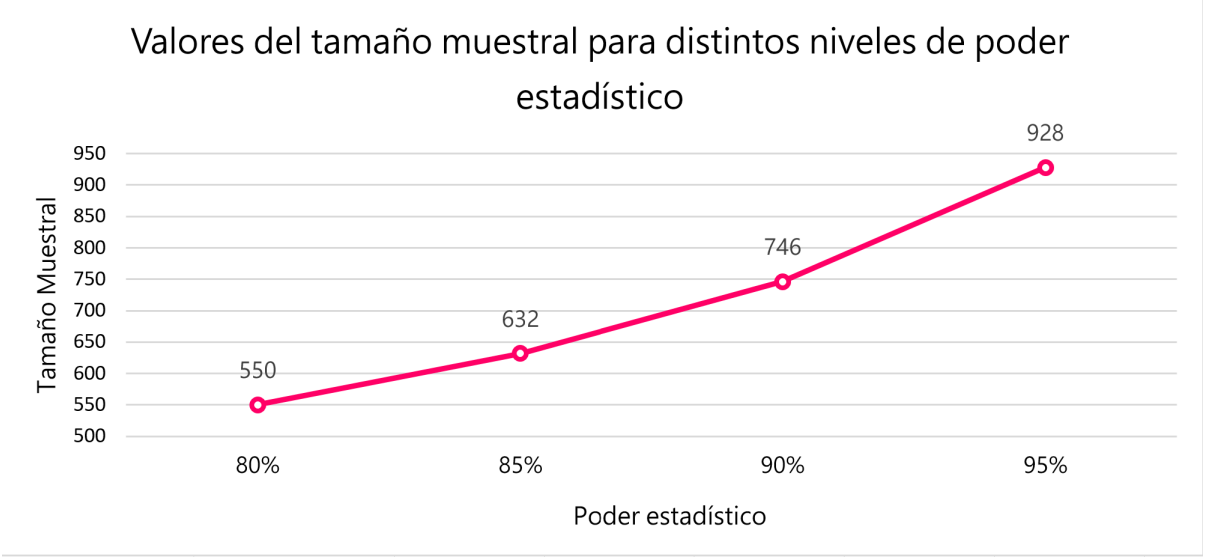


Figura 6.17: Gráfico de línea que muestra el tamaño muestral necesario para conseguir distintos niveles fijados para el poder estadístico.

Si bien se tienen los tamaños muestrales, para definir el poder estadístico que se utilizará en el experimento depende de los clientes activos promedio que hay mensualmente. Para hacer este cálculo, se extrajo el último año móvil (desde septiembre de 2020 hasta agosto de 2021), donde se obtuvo que en promedio se tiene un set de **420** clientes activos mensuales y una cantidad mensual promedio de **81** clientes activos que pasan a estar dentro del universo de clientes identificados como fuga. Si bien se tienen 420 clientes activos, de ese total se debe ejecutar el modelo de clasificación, el cual en su validación obtuvo una tasa de identificación de fugados de un 56.7 %. Por lo tanto, los clientes activos que se considerarían en el experimento serían **238 clientes**. En base a la Figura 6.17 y al número de clientes recién mostrado, el valor de poder estadístico que hace sentido fijar es de un **80 %**. Esta decisión se toma en

[8] El programa estadístico G*power de puede encontrar en la siguiente dirección: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

relación al tiempo que demoraría recolectar un total de **550** clientes activos, el cual sería de cinco meses en total.

En resumen, el experimento se realizaría en base a un test de proporciones de una cola, fijando un nivel de confianza de un 95 % y un poder estadístico de un **80 %**. Según los valores calculados, se necesita un total aproximado de **550 clientes activos** para que el experimento pueda capturar un tamaño del efecto (diferencia en las desviaciones estándar de los grupos) de un **7.6 %**. Además, según el promedio de clientes activos mensuales del último año móvil aplicando el modelo de clasificación (238 clientes) y cantidad mensual promedio de clientes activos identificados nuevos (81 clientes), se requieren **cinco meses** de experimento para poder tener conclusiones que estadísticamente tengan validez.

6.6. Análisis Financiero

Finalmente, se tiene uno de los más importantes análisis del trabajo de título, el cual trata del balance financiero que implica hacer la implementación de lo desarrollado a lo largo de este documento. La finalidad de este apartado es situarse en distintos escenarios dependiendo del tiempo que requiere y éxito que se alcance en el experimento realizado. De tal forma que se pueda ver una estimación del costo-beneficio que se tendría al aplicar un modelo de predicción de fuga de clientes y sus respectivas medidas comerciales para incentivar a los clientes etiquetados a reoperar.

6.6.1. Datos utilizados y supuestos

Lo primero que se debe mencionar antes de realizar los cálculos financieros es fijar la ventana de tiempo que se utilizará para obtener los datos con los que se trabajará. En particular, los datos que se usaron para el análisis financiero pertenecen al último año móvil de la misma base de datos con la que se evaluaron los modelos de clasificación, o sea desde septiembre de 2020 hasta agosto de 2021. Sin embargo, en este caso se incorporaron columnas que tienen relación con las tasas cobradas, los costos de operación de cada factura, segmentación por tamaño y riesgo que tiene la empresa de los clientes.

Tal como se dijo, se incorporaron nuevas columnas con información valiosa para hacer un análisis financiero. No obstante, para entender de una forma más profunda la información incluida se enlistarán, al igual que en la sección 6.4, los instrumentos que tiene Chita para balancear los precios. Estos son los siguientes:

- Tasa de Interés asociada a la factura
- Comisión asociada a la factura
- Porcentaje de Anticipo

En primer lugar, se tiene a la Tasa de Interés, que es el instrumento más dinámico que maneja la empresa, pues depende de factores como lo es el monto de la factura, segmentación por tamaño interno de la empresa y clasificación de riesgo. Por la confidencialidad y sensibilidad de esta información no se puede mostrar con detalle todos los valores y condiciones que fija la empresa para el cálculo. Sin embargo, se tomarán algunos ejemplos de tasas de interés

como comportamiento promedio para poder hacer la simulación.

En segundo lugar, se tiene a la Comisión que depende exclusivamente del monto de la factura. Por razones similares a lo anterior no se puede mostrar la tabla de precios, pero de igual forma será un instrumento que mostrará el comportamiento promedio al momento del análisis. Por último, se tiene el porcentaje del Anticipo. Para éste instrumento, se el promedio de la ventana temporal utilizada junto al valor fijado para la aplicación de la oferta propuesta.

Adicionalmente, para realizar el análisis financiero se debe estimar el costo de contacto de los clientes que recibirán la oferta propuesta en 6.4. Para esto, se toma como supuesto $600.000 de sueldo que recibe un ejecutivo comercial, quien es el encargado del contacto con los clientes vía correo y/o teléfono. Considerando datos entregados por el Área Comercial, se considerará que el envío de un correo toma en promedio 3 minutos y un llamado toma en promedio 5 min. Así, para calcular el costo de contacto se calculará un proporcional equivalente en dinero (pesos chilenos) del tiempo utilizado por ejecutivos comerciales en contactar a los clientes del experimento. Considerando 20 días hábiles y 8 horas diarias trabajadas por los ejecutivos comerciales, el costo por contactar a 1 cliente perteneciente al grupo de tratamiento del experimento es de $313 por una llamada telefónica y $188 por enviar un correo.

6.6.2. Simulación de escenarios

La simulación del análisis financiero es una etapa que permite visualizar de una manera muy parecida a la realidad de lo que podría ocurrir en términos monetarios sobre los distintos escenarios según las tasas de fuga y retención. Como se dijo, se tomaron valores promedio de los clientes activos según la segmentación de tamaño de la empresa (menos detallada que la del modelo de clasificación), riesgo de pago del cliente y tamaño de la factura (se categoriza según el monto, ver Anexo C.1). Para poder realizar un análisis financiero, se necesario dividir el trabajo en dos etapas, una para simular el efecto monetario de realizar el experimento (corto plazo) y otra para simular la implementación a nivel de negocio de la estrategia de oferta en un plazo de 12 meses (mediano plazo).

6.6.2.1. Simulación diseño experimental

Para poder realizar esta simulación es fundamental tener conocimiento sobre algunos supuestos que se tomaron para los cálculos. Tal como se dijo en la sección 6.5., Se consideró una base promedio de clientes activos (420) que irán incrementando mensualmente según la cantidad de clientes activos identificados nuevos (81), alcanzando en 5 meses un total de **562 clientes activos**. Adicionalmente, según los datos del último año móvil de datos, se tiene que en promedio opera el **73 %** de los clientes activos, con **2 operaciones** mensuales y **2 facturas** por operación. Por parte de la utilidad por factura, se calculó un promedio ponderado considerando una distribución de distintos tamaños de empresas y categorías de facturas según su monto (Ver Anexo C.1). Los distintos escenarios estarán basados en los valores para la tasa de fuga (10 %, 15 %, 20 % y 25 %) y para la tasa de retención (5 %, 10 % y 15 %).

Con el fin de calcular el beneficio neto de implementación del proyecto, se utilizarán las siguientes ecuaciones para cada comparación, sin experimento (SE) y con experimento (CE):

$$\Pi_{se} = N \cdot \alpha \cdot f \cdot c \cdot ut\_se \cdot (1 - \beta) \tag{6.1}$$

$$\Pi_{ce} = N \cdot p \cdot \alpha \cdot f \cdot c \cdot [(1 - \beta)ut\_se + (1 + \gamma - \beta)ut\_ce] \tag{6.2}$$

Donde N es la cantidad de clientes activos que participan en el experimento, α la proporción de clientes que en promedio operan mensualmente, p la proporción de clientes que pertenecen al grupo de tratamiento (50 %), f la cantidad promedio de facturas operadas mensualmente, c la cantidad de operaciones promedio realizadas mensualmente, β la tasa de fuga, γ la tasa de retención, ut_se la utilidad promedio sin experimento y ut_ce la utilidad promedio con experimento.

En base a las fórmulas 6.1 y 6.2, se hizo el cálculo financiero de la simulación del diseño experimental, resultando lo mostrado en la Tabla 6.10.

Tabla 6.10: Tabla que muestra los distintos escenarios para el análisis financiero para la implementación del experimento diseñado.

Escenarios	**Utilidad SE**	**Utilidad CE**	**Dif. Utilidad**	**Clientes SE**	**Clientes CE**	**Dif. Clientes**	β	γ
E1	$ 74.186.761	$ 64.536.014	▼ − 13.0 %	506	520	▲2.8 %	10 %	5 %
E2	$ 70.065.274	$ 61.030.921	▼ − 12.9 %	478	492	▲2.9 %	15 %	5 %
E3	$ 65.943.788	$ 57.525.829	▼ − 12.8 %	450	464	▲3.1 %	20 %	5 %
E4	$ 61.822.302	$ 54.020.736	▼ − 12.6 %	422	436	▲3.3 %	25 %	5 %
E5	$ 74.186.761	$ 65.980.363	▼ − 11.1 %	506	534	▲5.6 %	10 %	10 %
E6	$ 70.065.274	$ 62.475.270	▼ − 10.8 %	478	506	▲5.9 %	15 %	10 %
E7	$ 65.943.788	$ 58.970.178	▼ − 10.6 %	450	478	▲6.3 %	20 %	10 %
E8	$ 61.822.302	$ 55.465.085	▼ − 10.3 %	422	450	▲6.7 %	25 %	10 %
E9	$ 74.186.761	$ 67.424.712	▼ − 9.1 %	506	548	▲8.3 %	10 %	15 %
E10	$ 70.065.274	$ 63.919.619	▼ − 8.8 %	478	520	▲8.8 %	15 %	15 %
E11	$ 65.943.788	$ 60.414.527	▼ − 8.4 %	450	492	▲9.4 %	20 %	15 %
E12	$ 61.822.302	$ 56.909.434	▼ − 7.9 %	422	464	▲10.0 %	25 %	15 %

Según estos resultados, se puede apreciar que en todos los escenarios propuestos para la realización del experimento, en términos de la utilidad, conlleva en promedio un beneficio neto de un **10.7 %** menor que no hacerlo, equivalente a **$ 7.268.443**. Mientras que para los clientes ocurre lo contrario. En promedio se obtuvo que al realizar el experimento se mantienen activos un **6.1 %** más de clientes que al no hacer el experimento. Este comportamiento de los números es interesante, pues es coherente que la utilidad sea menor, ya que se está incurriendo en una inversión aplicada a la estrategia de oferta, resultando mayor el gasto que el beneficio neto. Sin embargo, la cantidad de clientes es mayor, lo que se condice con el objetivo de la estrategia de oferta, es decir, poder retener a clientes posibles de fuga y que en total se tenga una mayor cantidad de clientes activos.

Si bien estos resultados, a nivel de las utilidades no parece rentable para llevar a cabo, como se mencionó al comienzo de esta sección, esta simulación está orientada solo a la realización del experimento (corto plazo). Para obtener resultados más robustos se debe recurrir a la segunda etapa de la sección de simulación, donde se profundizará la implementación de la estrategia de oferta y se podrá simular el efecto que ésta produce en las utilidades y cantidad de clientes acumulados en un periodo de 12 meses.

6.6.2.2. Simulación implementación

Tal como se pudo apreciar, los resultados que se obtuvieron desde la simulación del experimento son confusos hasta el momento en su interpretación, puesto que no se cumplen todos los objetivos (obtener una mayor utilidad y mantener activos a más clientes). Debido a esto, se recurrió a una simulación que ayudara a complementar lo realizado en la sección 6.6.2.1. de una manera más robusta y a mediano plazo (12 meses).

Al igual que para la simulación anterior, es importante revisar los supuestos que se utilizaron para poder hacer los cálculos financieros. Algunos supuestos son compartidos con lo realizado anteriormente como la cantidad de operaciones promedio (2), la cantidad de facturas promedio (2), la proporción de clientes activos que en promedio operan (73 %) y las utilidades promedio para clientes sin oferta como para clientes con oferta. Adicionalmente, se incorpora una nueva cantidad base de clientes activos alcanzando un total promedio de 420 clientes, una cantidad de clientes nuevos promedio que ingresan mensualmente (86 clientes nuevos), una tasa de identificación de fuga (usando el modelo de clasificación) de un 56.7 %, una tasa de fuga para clientes sin oferta de un 18 % y otra para clientes identificados de un 25 % y la proporción de clientes que se utilizarán del grupo de clientes identificados como posible fuga (20 %).

En el caso de la simulación para la implementación de la estrategia de oferta se construyeron los escenarios posibles en relación con la variación de la tasa de retención, la cual toma los valores de 5 %, 8 %, 10 % y 15 %. Cada escenario busca comparar en un mismo periodo la utilidad y cantidad de clientes acumulados resultante, aplicando o no una oferta a un grupo de clientes identificados. Para poder realizar el cálculo de las métricas mencionadas, se parte de la base de **420 clientes activos**, donde con el modelo en promedio se identifica al 56.7 % de los clientes como posibles a fugarse. Sin embargo, aplicar una oferta a un grupo de 238 clientes implica un costo importante para la empresa. Por lo tanto, se propone utilizar al 20 % de clientes identificados con la mayor probabilidad proporcionada por el modelo de clasificación, alcanzando un total de **48 clientes**, a los cuales se les aplicará la oferta según las tasas de fuga y retención.

Dada toda la información y al igual que para la simulación anterior, se definen las ecuaciones 6.3, 6.4, 6.5 y 6.6, la cuales apoyaron el cálculo del beneficio neto y clientes activos acumulados.

$$\Pi_{so} = \sum_{i=1}^{12} N_i(1-\beta_{so})\alpha\cdot f\cdot c\cdot ut\_so + n_i\cdot f\cdot c\cdot ut\_so \tag{6.3}$$

$$\Pi_{co}^{1} = N_1(1-t_{id}\cdot l)\alpha\cdot f\cdot c\cdot ut\_so + n_1\cdot f\cdot c\cdot ut\_so + N_1\cdot t_{id}\cdot l(1+\gamma-\beta_{co})f\cdot c\cdot ut\_co \tag{6.4}$$

$$\Pi_{co}^{2-12} = \sum_{i=2}^{12} N_i(1-t_{id}\cdot l)\alpha\cdot f\cdot c\cdot ut\_so + n_i\cdot f\cdot c\cdot ut\_so + N_i\cdot t_{id}\cdot l(1+\gamma-\beta_{co})f\cdot c\cdot ut\_co \tag{6.5}$$

$$\Pi_{co} = \Pi_{co}^{1} + \Pi_{co}^{2-12} \tag{6.6}$$

Donde N_i es la cantidad de clientes activos por cada mes, n_i es la cantidad de clientes nuevos que se agregan cada mes, α la proporción de clientes que en promedio operan mensualmente, t_{id} la tasa de clientes identificados por el modelo de clasificación, l la proporción de clientes identificados que se eligen para aplicar la estrategia de oferta, f la cantidad promedio de facturas operadas mensualmente, c la cantidad de operaciones promedio realizadas mensualmente, β_{so} la tasa de fuga para clientes que no se aplica estrategia, β_{co} la tasa de fuga para clientes que se aplica estrategia, γ la tasa de retención, ut_so la utilidad promedio sin oferta y ut_co la utilidad promedio con oferta.

Dada la información de los supuestos y las ecuaciones planteadas para realizar los cálculos financieros, se desprenden los siguientes resultados.

Tabla 6.11: Tabla que muestra los distintos escenarios para el análisis financiero para la implementación del experimento diseñado.

Escenarios	**Utilidad SO**	**Utilidad CO**	**Dif. Utilidad**	**Clientes SO**	**Clientes CO**	**Dif. Clientes**	γ
E1	$ 861.867.111	$ 862.584.317	▲0.1 %	472	480	▲1.7 %	5 %
E2	$ 861.867.111	$ 880.686.239	▲2.2 %	472	497	▲5.2 %	8 %
E3	$ 861.867.111	$ 893.084.557	▲3.6 %	472	509	▲7.7 %	10 %
E4	$ 861.867.111	$ 925.294.851	▲7.4 %	472	540	▲14.3 %	15 %

Según los resultados de la Tabla 6.11, se puede apreciar un comportamiento distinto de las utilidades respecto a la simulación del experimento. Para una tasa de retención del 5 %, los valores son muy similares, sin embargo, desde un 8 % para la retención en adelante, se observa una diferencia significativa. En términos porcentuales, se obtuvo en promedio un aumento de un **3.3 %** de las utilidades equivalente a **$ 28.545.380**. Asimismo, la cantidad de clientes activos acumulados que se mantienen al aplicar la estrategia de oferta es mayor a que no aplicarla, obteniendo un aumento promedio de un **7.2 %**.

A diferencia de la simulación anterior, se puede apreciar que se cumple el objetivo planteado detrás de la aplicación de la estrategia comercial de oferta. Esta estrategia busca principalmente mantener retenidos a clientes posibles de fuga, de manera que en total se tenga un número mayor de clientes activos, lo cual a mediano plazo (12 meses) genera mayor utilidades que no aplicar una estrategia proactiva de retención. Es por esta razón, que al realizar la simulación del experimento (que ya mostraba resultados favorables en relación a los clientes) y complementarla con una simulación más robusta que realiza un ejercicio más apegado a la realidad, se puede establecer que la estrategia de oferta es una potencial forma de combatir la fuga de clientes de manera proactiva.

Si bien se hizo un proceso que mostrara robustez para la simulación de escenarios financieros respecto a la aplicación de una estrategia de retención, hay muchos supuestos detrás que pueden hacer variar para bien o para mal los cálculos, lo cual evidentemente representa un riesgo involucrado. Los factores que mayor influencia pueden tener sobre las simulaciones son las tasas de fuga y retención, la proporción de clientes que operan en promedio mensualmente o la cantidad de clientes activos/nuevos, debido a variaciones que puedan tener en el tiempo. Estas son consideraciones importantes que deben tomarse en cuenta a la hora de una futura implementación, sin embargo, la base para poder realizar este proceso, está mostrada dentro de este trabajo de título.

Capítulo 7

Conclusiones

Luego de haber realizado todas las etapas planteadas y descritas en el trabajo de título, es necesario relevar la importancia de la metodología utilizada, pues con esta se logró organizar y planificar los distintos tiempos de aprendizaje y ejecución sobre el tema trabajado. Gracias a la metodología CRISP-DM se alcanzó un conocimiento importante sobre el rubro del Factoring, un rubro B2B que está en crecimiento, pero que se vio muy afectado por el contexto del COVID-19.

Fue fundamental ahondar en el problema encontrado ya que para la empresa es un dolor vigente y con una alta probabilidad seguirá siendo un tema a mejorar constantemente, ya que toda empresa necesita a sus clientes para subsistir. La fuga de clientes es un problema que se puede atacar de diferentes aristas, con distintas herramientas o técnicas. En particular, para este trabajo de título los esfuerzos se centraron en los clientes activos que mantiene la empresa. Para conocer más de este conjunto de clientes se hizo una búsqueda de fuentes de datos, de las cuales se encontró información valiosa del comportamiento y experiencia de los clientes. Sin embargo, se identificaron algunas deficiencias en la captura de datos. Por ejemplo, dentro de las bases disponibilizadas por la Gerencia Comercial no existía información demográfica de los clientes, para lo cual se recurrió a bases de datos públicas del Servicio de Impuestos Internos (SII). Otro ejemplo de las bases de datos es respecto a las variables de experiencia y su consistencia, puesto que se esperaba una mayor cantidad de puntos medidos y capturados en los datos de la empresa. Más aún, de las variables existentes se localizaron algunas que no tenían una consistencia a lo largo de la base de datos. Es por esta misma razón que se decidió trabajar con datos desde el año 2018 en adelante, ya que algunos de los datos del año 2017 estaban en construcción y existían muchos valores vacíos. No se espera tener este tipo de problemas al tratar temas como la fuga de clientes. Si bien, estas deficiencias fueron un desafío y, a la vez, parte de las decisiones que se tomaron a la hora de elegir las ventanas de tiempo o variables con las que se trabajaron, no fue un impedimento para construir una base de datos que mostrara el comportamiento de los clientes.

Así como fue un desafío la búsqueda de las bases de datos y su comprensión para el desarrollo de este trabajo, la preparación de la información también lo fue. El mayor reto que se presentó en la preparación de los datos para los modelos fue construir una base de datos que siguiera una línea conductora que contara la historia de los clientes y que permitiera unir la comprensión del negocio y el problema con los requerimientos que tienen los modelos de clasificación utilizados. En esta etapa se dedicó la mayor parte del tiempo debido a la

gran cantidad de iteraciones que se realizaron para llegar al producto deseado. De hecho, una vez que se tuvo la base de datos que se utilizaría con los modelos, ésta sufrió por lo menos 3 cambios grandes que permitió un mejor entendimiento del fenómeno de la fuga y cómo ligar ese conocimiento con la forma que los modelos de clasificación aprenden de los datos.

La etapa de modelamiento fue otra de las secciones a las que se le dedicó una gran parte del tiempo utilizado para el desarrollo del trabajo de título. Resulta evidente destinar un gran esfuerzo en conseguir modelos que clasifiquen de la mejor manera posible, ya que el proceso depende de diversos factores que deben estar bien ejecutados y basados en la teoría para que se cumpla el objetivo de ese apartado. En particular, en una fase temprana del modelamiento se obtuvieron resultados que tenía un comportamiento y magnitudes no tan coherentes con lo aprendido del negocio. Esto fue un punto de inflexión con el cual se apalancó el desarrollo de una mejor etapa de construcción de las bases de datos (tal como se mencionó anteriormente), lo cual permitió que el rendimiento de los modelos cobrara sentido de negocio.

Luego de muchas iteraciones, se encontró el modelo que cumplía con las mejores condiciones para cumplir con la finalidad del modelamiento, un Árbol de Clasificación. Este modelo cobró mucho sentido de negocio al momento de analizar la importancia de las variables utilizadas para su ejecución. Dentro de estas variables, la más importante fue la Recencia sobre la Antigüedad (R/A). Que ésta variable se la más importante tiene bastante coherencia con el trasfondo y lo que comunica el valor de R/A, pues es un ratio entre los días que han pasado desde la última operación realizada por un clientes sobre la antigüedad del mismo. De hecho, es la misma Recencia la que define directamente en su valor si un cliente está fugado o no. Es por esta razón que no se incluyó directamente y se le aplicó una transformación junto a la Antigüedad. El resto de las variables importantes dentro del modelo elegido también forman parte del hilo conductor que habla sobre el comportamiento y experiencia de los clientes, como lo es el Monto acumulado de las facturas operadas, los Días de mora promedio que tiene un cliente o las horas que pasan desde la cesión de una factura hasta que la empresa gira el dinero. De esta forma, sin poner atención a los resultados en el etiquetado, el modelo tiene un sentido de negocio que permite analizar la fuga.

Por su parte, los resultados que tienen que ver con el etiquetado del modelo de clasificación tienen un buen desempeño en término de sus métricas. El modelo elegido obtuvo en su fase de testeo un 76.3 % de exactitud (accuracy) en la identificación tanto de clientes activos como fugados y un valor de 31.8 % para el etiquetado de falsos positivos. A pesar de que el modelamiento es una tarea que contempla muchas etapas necesariamente bien ejecutadas, en base a sus resultados, como conclusión del modelo elegido, este cumplió con ser robusto en la clasificación de clientes según su estado tal como se pudo ver en la validación (donde incluso etiquetó mejor a clientes activos que fugados). Si bien se busca identificar clientes potenciales a fugarse, también es relevante para controlar mejor los gastos, tener una buena precisión en la identificación de clientes activos.

Con el modelo construido, validado y elegido, se pasó a otra oportunidad de negocio importante que fue proponer una medida comercial basada en una oferta respecto a los costos de servicio que percibe un cliente. Lo más complejo de poder hacer esta propuesta fue el acotado abanico de posibilidades donde recurrir para incentivar a los clientes a la reoperación. Como se comentó en el desarrollo, solo se contaba con la tasa de interés, comisión y porcentaje del

anticipo. En base a la comprensión del negocio se pudo identificar que el rubro del Factoring posee márgenes bajos respecto a otros rubros, por lo que es una difícil tarea hacer descuentos sin sacrificar parte de las utilidades. A pesar de todo lo mencionado, se logró construir una propuesta llamativa y que incentiva a que el cliente opere al menos dos veces más con la empresa.

Dado el contexto de la empresa y los tiempos necesarios para implementar mediciones de lo construido y mencionado anteriormente, se tomó la decisión de dejar planteado y diseñado el experimento que se debe ejecutar para poder medir el impacto de identificar a clientes potenciales a fugarse con proactividad. A pesar de quedar planteado y no ejecutarse, se definen las bases teóricas y su bajada concreta al caso de estudio donde se explicitan requerimientos importantes como lo es el tiempo de implementación. Para el diseño se necesitan alrededor de 550 clientes dentro del experimento para poder medir un efecto con validez estadística, lo que implica ejecutar el experimento durante cinco meses continuos. La duración recién mencionada puede ser un gran obstáculo a nivel comercial, pues es un largo periodo de tiempo. Sin embargo, para lograr medir un efecto se necesita una cantidad no menor de clientes y la empresa al ser del tipo B2B no posee un flujo tan abundante de clientes como en un rubro B2C como lo es el retail. Por lo tanto, se cumple con dejar todas las indicaciones y supuestos que se deben tener para poder hacer un experimento, implementación que dependerá exclusivamente de la empresa y sus prioridades.

Finalmente, con el objetivo de hacer más robusto en proceso de diseño de una eventual implementación del trabajo de título, se realizó un análisis financiero, con el cual se pudo comprobar lo mencionado respecto a los bajos márgenes del rubro del Factoring. Dado que existe un espacio acotado para incentivar a los clientes a través de medidas como descuentos o paquetes de descuentos, los resultados del análisis respaldan la teoría de que para atraer a que los clientes operen nuevamente se deben sacrificar las utilidades en el corto plazo. Para apoyar esto se realizó una simulación de implementación del diseño experimental, con la cual se pudo evidenciar que en términos de las utilidades percibidas por la empresa, en promedio disminuyen en un 10.7 %, equivalente a $ 7.268.443. Mientras a nivel de clientes, la simulación del experimento cumplió su objetivo, alcanzando una mayor cantidad de clientes que al no aplicar el experimento de un 6.1 % (28 clientes). Sin embargo, no es suficiente respaldar solo el corto plazo, más aún cuando existe una pérdida de utilidades dada por una inversión en una estrategia de retención. Por esta razón, se complementó el trabajo realizado con la simulación del experimento con una simulación de una eventual implementación de la estrategia de retención en el mediano plazo (12 meses). De esta simulación de obtuvo que al aplicar una oferta a un grupo acotado de clientes identificados (20 % de los clientes etiquetados por el modelo con mayor probabilidad de fuga) junto con la operación promedio mensual de la base de clientes activos de Chita, se obtuvo un aumento de un 3.3 % ($ 28.545.380) y un 7.2 % (34) para las utilidades acumuladas percibidas y la cantidad de clientes activos acumulados, respectivamente.

Es fundamental destacar que los resultados de ambas simulaciones complementadas entre sí son favorables y plantean un escenario potencial para implementar lo planteado en el trabajo de título, ya que los resultados muestran una dirección alineada con los objetivos planteados en este documento. Es decir, identificar proactivamente a clientes potenciales de fuga, a los cuales mediante el proceso descrito, se incentive a operar con la empresa Chita de

modo que el flujo de clientes recurrentes y activos aumente, combatiendo de esta manera la tasa de fuga. Sin embargo, aún con resultados favorables para las simulaciones, es importante tener en cuenta que para los cálculos se aplicaron supuestos que pueden presentar variabilidad en las métricas calculadas como lo son las tasa de fuga y retención, la proporción de clientes activos que operan mensualmente o la cantidad de clientes activos/nuevos que se tienen de base. Si bien se trabajó para que los resultados empíricos sean lo más parecido a lo que pueda suceder en la realidad, se deja el espacio para que en una potencial implementación de este trabajo se pueda discutir el riesgo en base a un cálculo más preciso de los factores que inciden en el proceso.

7.1. Trabajos futuros

Con la finalidad de poder realizar una buena implementación del trabajo desarrollado en este documento, se dejan propuestos posibles cambios o mejoras al proceso de manera que desde la captura de datos hasta etapas más prácticas como el diseño experimental o derechamente la implementación de la estrategia proactiva de retención.

1. Conseguir y acoplar a las bases de datos utilizadas datos que representen otro tipo de segmentaciones de clientes como lo son las campañas de marketing digital que Chita utiliza para captar clientes.

2. Buscar más puntos de interacción con el cliente en las bases de datos de la empresa de modo que la experiencia del cliente esté presente en los datos utilizados por los modelos, para así tener una mayor y cercana sensibilidad de la satisfacción del cliente con el servicio.

3. Incorporar técnicas más innovadoras como programas de retención o fidelización de clientes a modo de tener un sistema más robusto y con incentivos más llamativos que apoyen la proactividad en la detección de clientes potenciales a fugarse en base a los modelos de clasificación.

4. Tener en consideración los supuestos de las simulaciones y considerar más variaciones y dinamismo, con datos actualizados al momento de implementar para mitigar los riesgos de implementación a corto, mediano y largo plazo.

Bibliografía

[1] Finnovating News – Conectando el ecosistema FinTech global. Definición de Fintech, 2017, https://www.finnovating.com/news/definicion-de-fintech/.

[2] ¿Qué es el Factoring? CMF. CMF, 2017, https://www.cmfeduca.cl/educa/600/w3-article-27145.html.

[3] SII | Tamaño de empresas. SII | Servicio de Impuestos Internos, 2018, https://www.sii.cl/sobre_el_sii/.

[4] Economía chilena se desploma en 2020 y sufre su peor caída en casi 40 años. América Economía, 2021, https://www.americaeconomia.com/economia-mercados/comercio/economia-chilena-se-desploma-en-2020-y-sufre-su-peor-caida-en-casi-40.

[5] Chile Panorama general. World Bank, 2021, https://www.bancomundial.org/es/country/chile/overview.

[6] CMF, Ley Nº 18.045 del Mercado de Valores. 2017, https://www.cmfchile.cl/portal/principal/613/w3-article-806.html.

[7] Efectos del Covid en la actividad de las Empresas en Chile. Comisión Nacional de Productividad, 2020.

[8] A new feature set with new window techniques for customer churn prediction in landline telecommunications. Expert Systems with Applications. Huang, B., Kechadi, T. M., Buckley, B., Kiernan, G., Keogh, E., & Rashid, T., 2017.

[9] A Self-Learning Text (Statistics for Biology and Health) (English Edition) (3.a ed.). Kleinbaum, D. G., 2010.

[10] Journal of Marriage and the Family. DeMaris, A., 1995.

[11] Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday: 109 (Softcover Reprint of the Original 1st 1996 ed.). Rieder, H., 1995.

[12] Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner (2nd ed.). Shmueli, G., Patel, N. R., & Bruce, P. C., 2010.

[13] Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R., 2020.

[14] Evaluating classification accuracy for modern learning approaches. CMLi, J., Gao, M., & D'Agostino, R.F, 2017.

[15] Comparison of logistic regression model and classification tree: An application to post-

partum depression data. CAMDEVIREN, H., YAZICI, A., AKKUS, Z., BUGDAYCI, R., & SUNGUR, M., 2007.

[16] Estudio comparativo de metodologías para minería de datos. Moine, J. M., Haedo, A., Gordillo, S., 2011.

[17] G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A., 2007.

Anexos

Anexo A. Caracterización de la empresa

Tabla A.1: Tabla que muestra la variación de las colocaciones desde el año 2017 al 2020. Fuente: Elaboración propia.

Año	Colocaciones
2017	-
2018	141 %
2019	76 %
2020	-13 %

Anexo B. Desarrollo metodológico

B.1. Análisis exploratorio

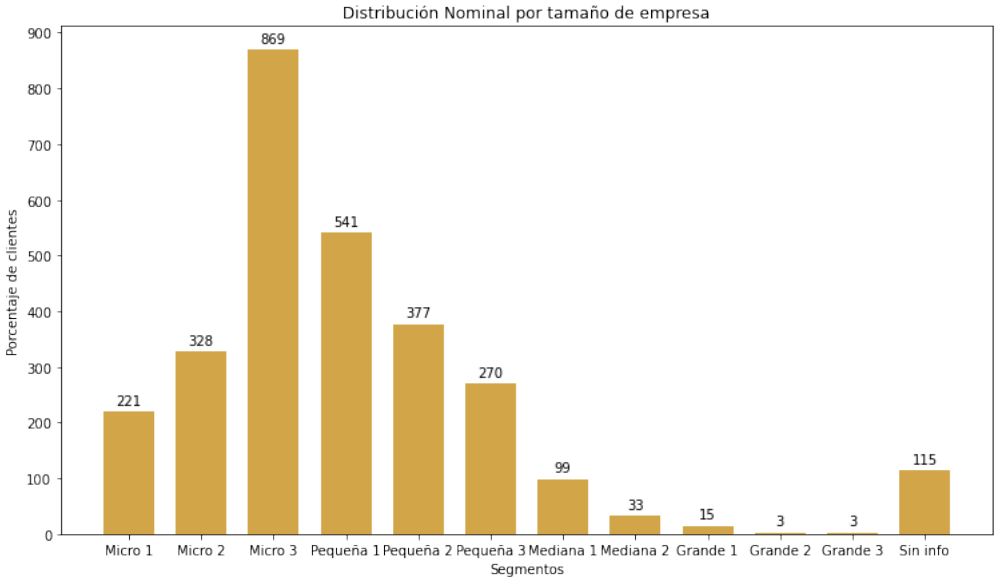


Figura B.1: Gráfico de barras que muestra la distribución nominal de clientes según tamaño. Fuente: Elaboración propia.

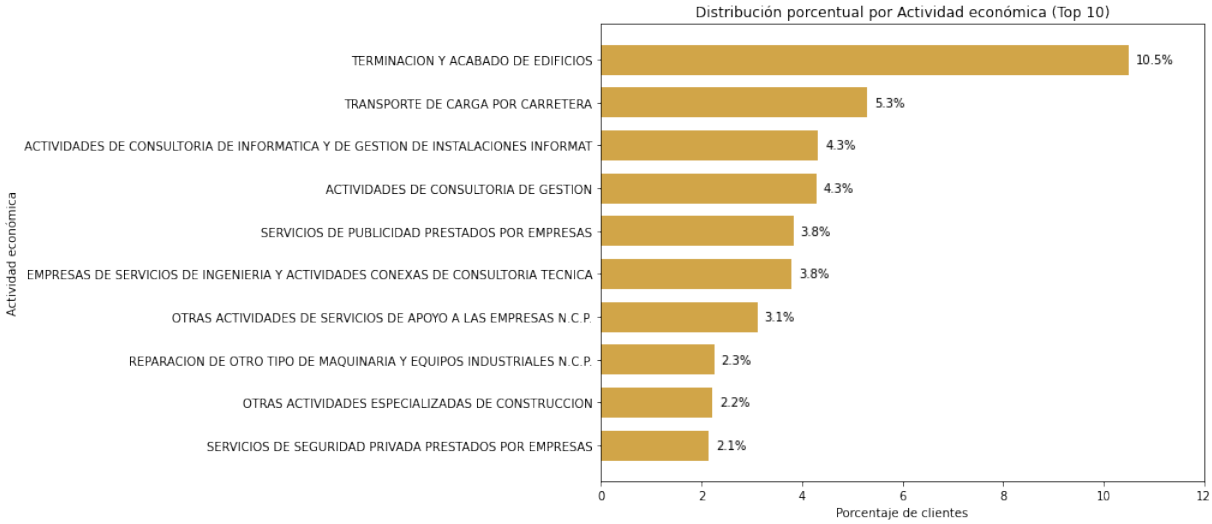


Figura B.2: Gráfico de barras que muestra la distribución porcentual de los clientes según las Top 10 actividades económicas. Fuente: Elaboración propia.

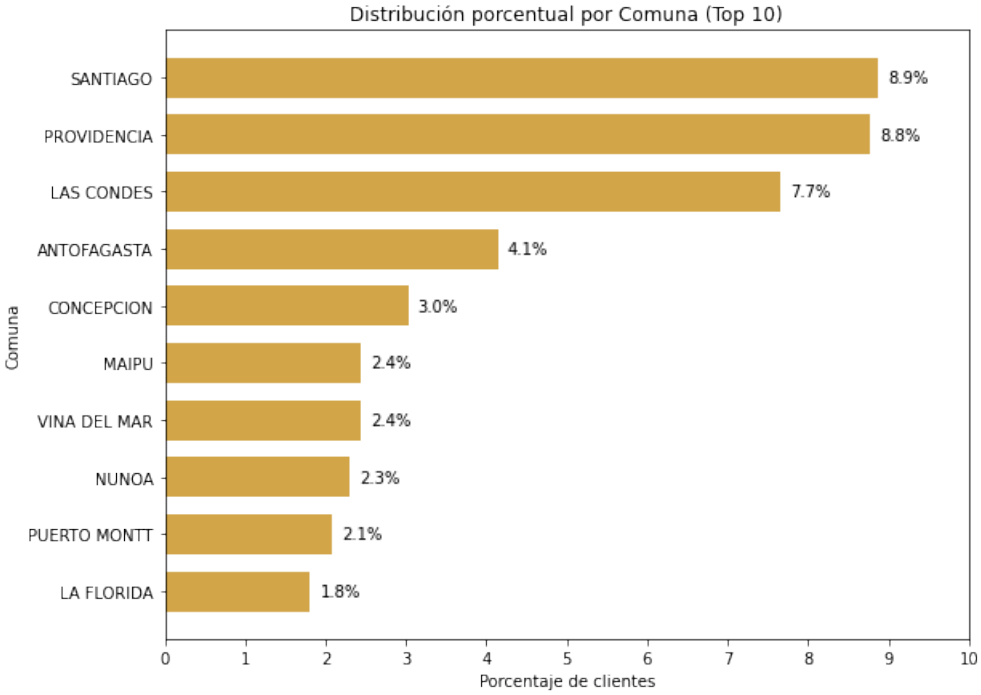


Figura B.3: Gráfico de barras que muestra la distribución del Top10 de comunas con mayor participación de clientes. Fuente: Elaboración propia.

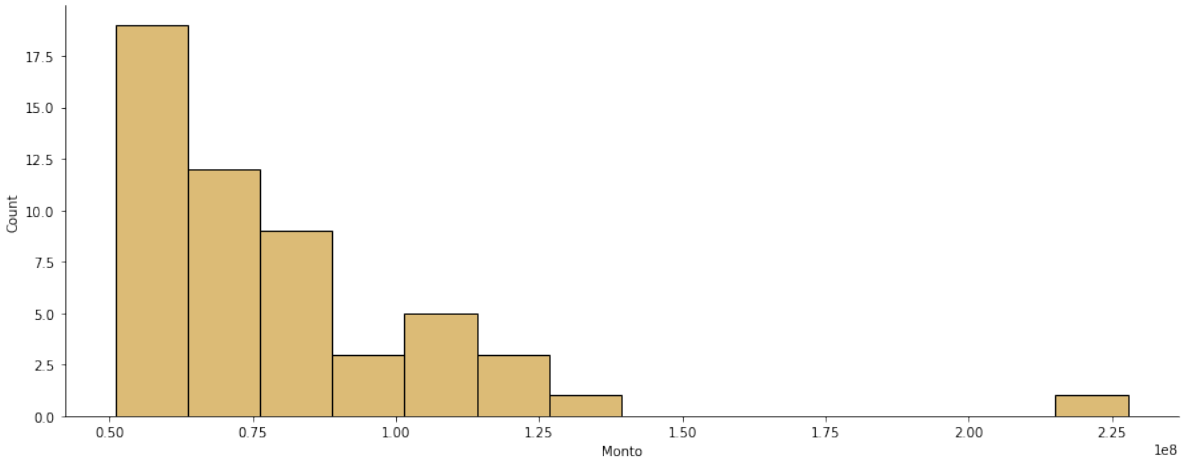


Figura B.4: Histograma que muestra la distribución del Monto de las facturas. Fuente: Elaboración propia.

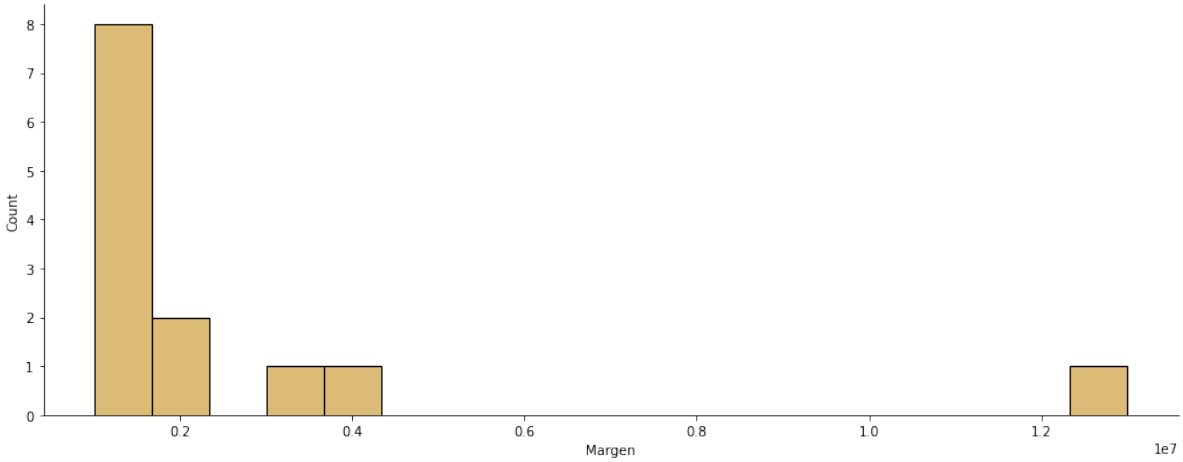


Figura B.5: Histograma que muestra la distribución del Margen de las facturas. Fuente: Elaboración propia.

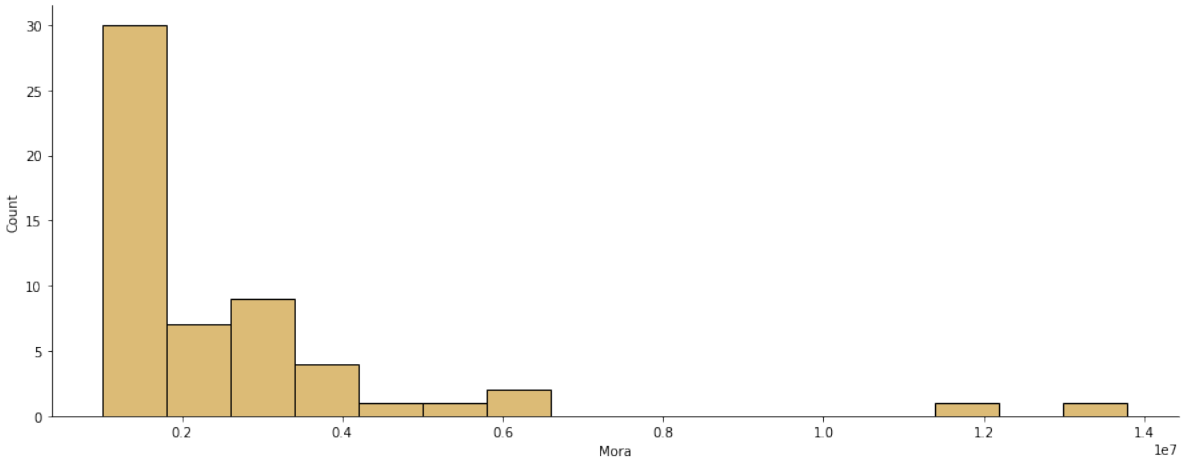


Figura B.6: Histograma que muestra la distribución de la Mora de las facturas. Fuente: Elaboración propia.

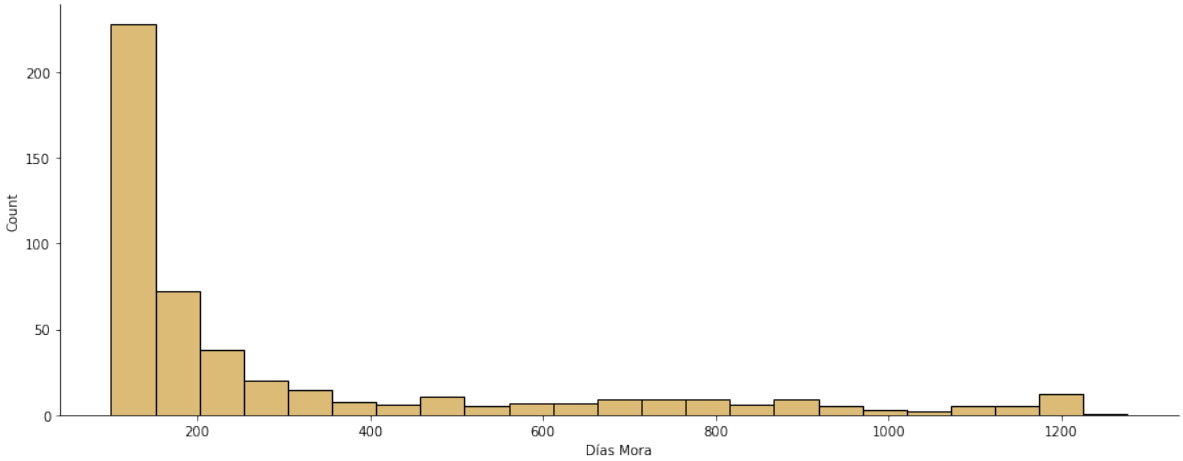


Figura B.7: Histograma que muestra la distribución de los Días de mora de las facturas. Fuente: Elaboración propia.

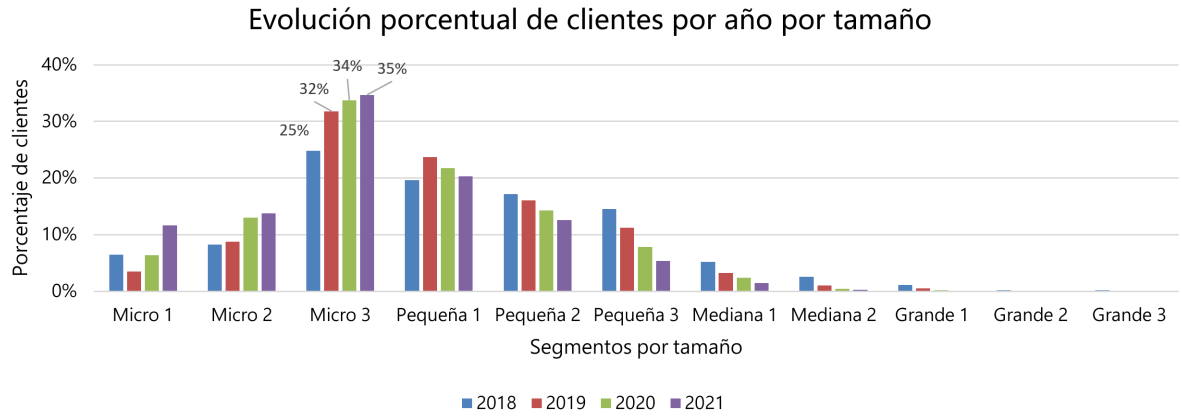


Figura B.8: Gráfico de barras que muestra la evolución porcentual de clientes por tamaño desde 2018 hasta agosto de 2021. Fuente: Elaboración propia.

B.2. Modelamiento

Tabla B.1: Tabla que muestra las métricas principales de todas las combinaciones de modelos probadas para la Regresión Logística, tomando el método de *Wrapping* como base.

Variables	**Notación Científica**
Monto	7.12E-10
Dias mora	4.74E-03
Frecuencia	-5.53E-03
Recesiones	3.2E-02
Hora giro	-3.8E-03
R/A	2.6
Mora	6.0E-01
Tamaño_Grande 2	1.4E-01
Tamaño_Grande 3	0
Tamaño_Mediana 1	1.3E-01
Tamaño_Mediana 2	3.1E-01
Tamaño_Micro 1	-9.1E-02
Tamaño_Micro 2	-8.3E-02
Tamaño_Micro 3	3.2E-02
Tamaño_Pequeña 1	4.7E-03
Tamaño_Pequeña 2	1.0E-02
Tamaño_Pequeña 3	2.0E-01
Rubro_ACTIVIDADES DE ALOJAMIENTO Y DE SERVICIO DE COMIDAS	-1.5E-01
Rubro_ACTIVIDADES DE ATENCION DE LA SALUD HUMANA Y DE ASISTENCIA SOCIAL	-5.7E-01
Rubro_ACTIVIDADES DE SERVICIOS ADMINISTRATIVOS Y DE APOYO	-1.3E-01
Rubro_ACTIVIDADES FINANCIERAS Y DE SEGUROS	0
Rubro_ACTIVIDADES INMOBILIARIAS	-2.8E-01
Rubro_ACTIVIDADES PROFESIONALES, CIENTIFICAS Y TECNICAS	-2.8E-02
Rubro_AGRICULTURA, GANADERIA, SILVICULTURA Y PESCA	5.4E-02
Rubro_COMERCIO AL POR MAYOR Y AL POR MENOR	1.9E-02
Rubro_CONSTRUCCION	4.8E-02
Rubro_ENSEÑANZA	3.2E-02
Rubro_EXPLOTACION DE MINAS Y CANTERAS	4.0E-01
Rubro_INDUSTRIA MANUFACTURERA	-5.2E-02
Rubro_INFORMACION Y COMUNICACIONES	9.1E-02
Rubro_OTRAS ACTIVIDADES DE SERVICIOS	2.3E-01
Rubro_SUMINISTRO DE AGUA	-1.5E-01
Rubro_SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR Y AIRE ACONDICIONADO	1.3E-01
Rubro_TRANSPORTE Y ALMACENAMIENTO	-6.2E-02

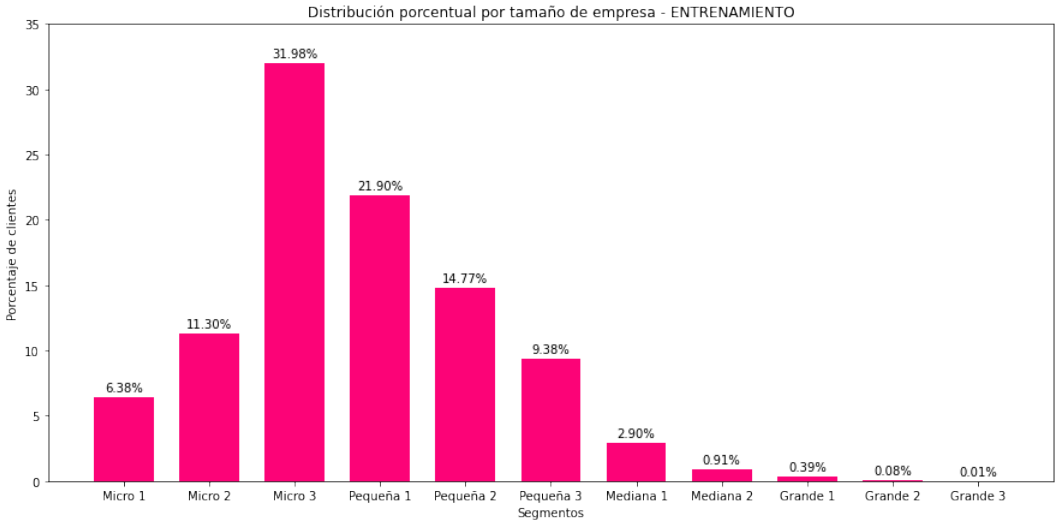


Figura B.9: Gráfico de barras que muestra la distribución porcentual por tamaño de las observaciones de la base de entrenamiento. Fuente: Elaboración propia.

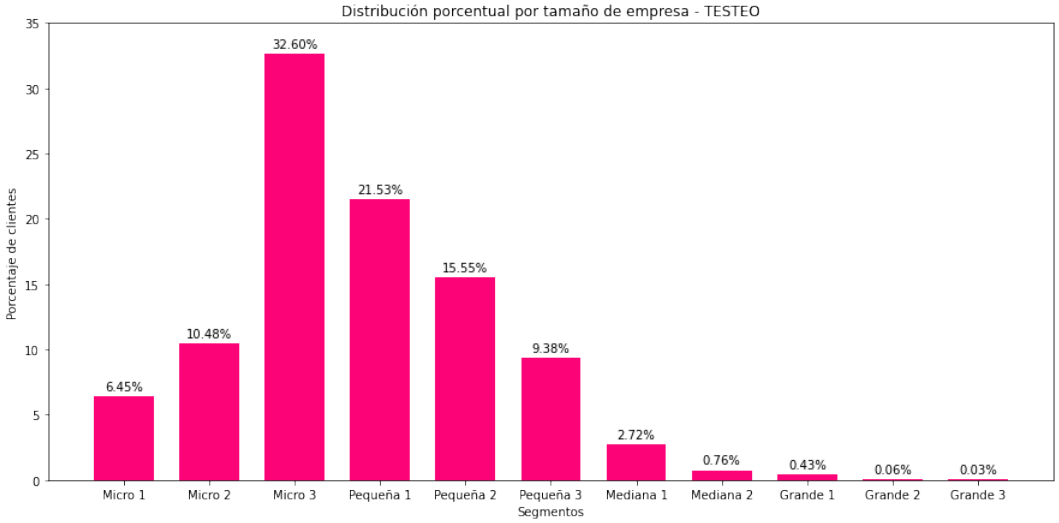


Figura B.10: Gráfico de barras que muestra la distribución porcentual por tamaño de las observaciones de la base de testeo. Fuente: Elaboración propia.

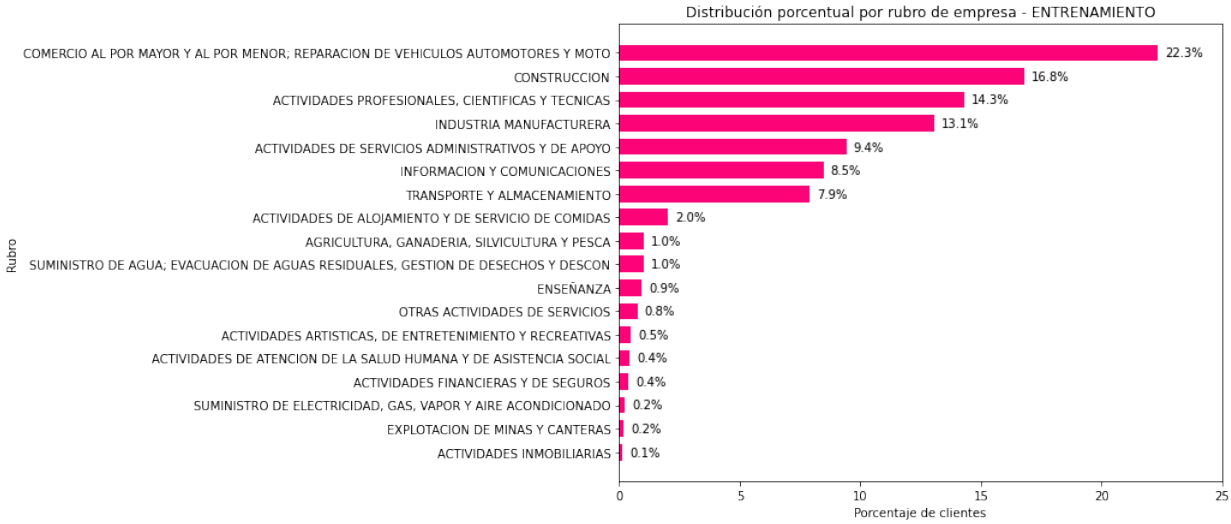


Figura B.11: Gráfico de barras que muestra la distribución porcentual por rubro de las observaciones de la base de entrenamiento. Fuente: Elaboración propia.

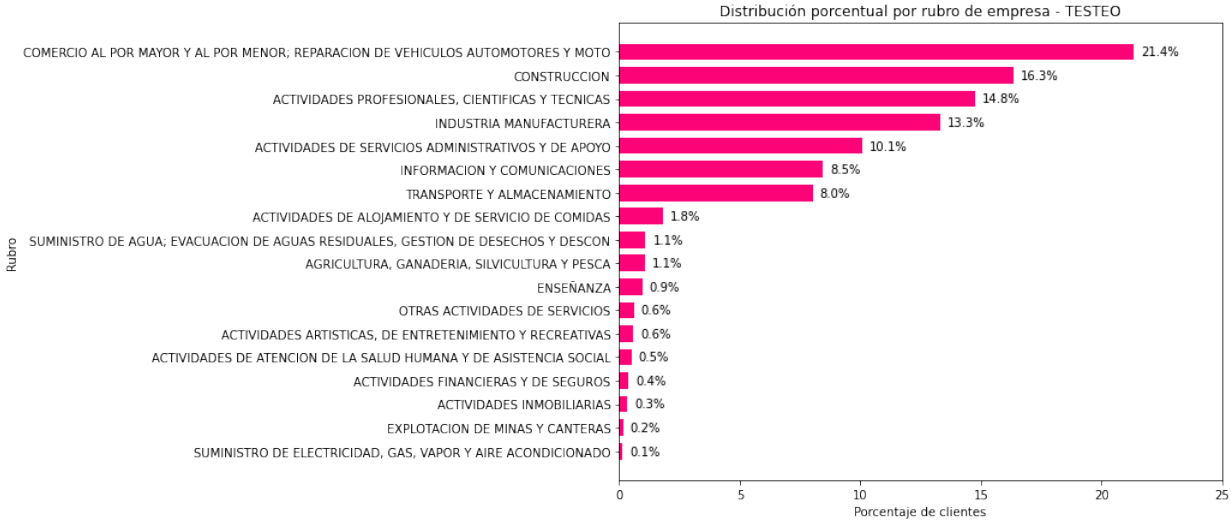


Figura B.12: Gráfico de barras que muestra la distribución porcentual por rubro de las observaciones de la base de testeo. Fuente: Elaboración propia.

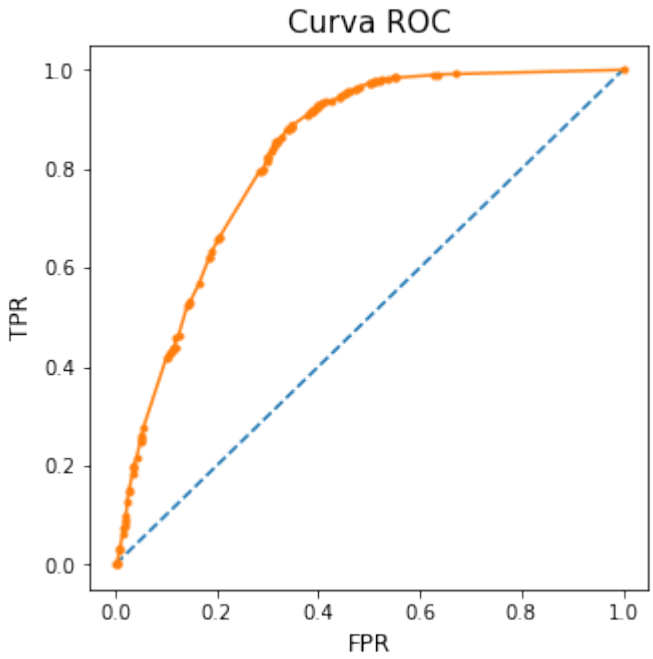


Figura B.13: Gráfico de barra que muestra la curva ROC del modelo de clasificación elegido (*Wrapping + Tamaño*). Fuente: Elaboración propia.

Anexo C. Diseño experimental

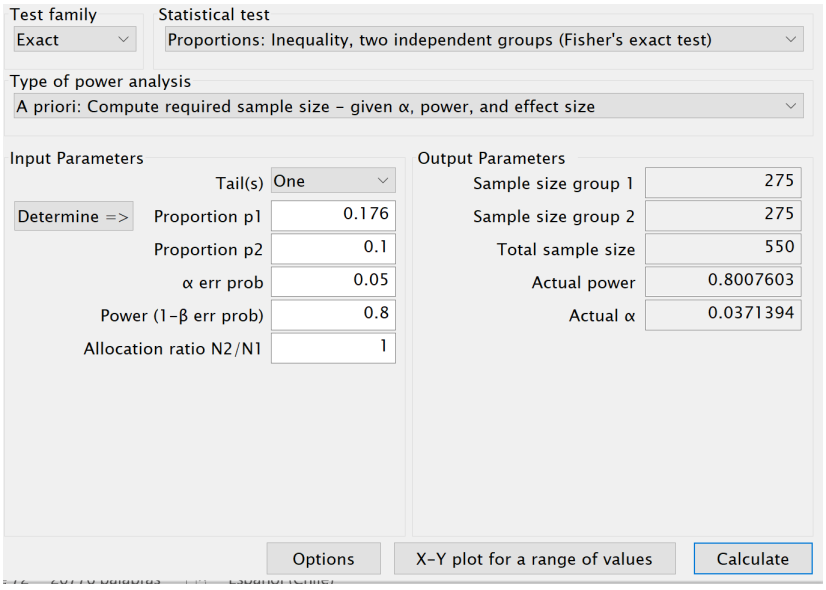


Figura C.1: Imagen que muestra un ejemplo del cálculo del tamaño muestral para un poder estadístico del 80 % en el programa *G*power*. Fuente: Elaboración propia.

C.1. Simulación de escenarios

Tabla C.1: Tabla que muestra la categorización de las facturas por tamaño según su monto. Fuente: Elaboración propia.

Tamaño factura	**Monto capturado**
Mega	[10.000.000 - 1.000.000.000]
Grande	[5.000.000 – 10.000.000)
Mediana	[1.500.000 – 5.000000)
Pequeña	[500.000 – 1.500.000)
Enana	[250.000 – 500.000)
Micro	[15.000 – 250.000)