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

DISEÑO DE MODELOS ECONOMÉTRICOS PARA LA ESTIMACIÓN DE ESTADOS FINANCIEROS DE MICROEMPRESAS QUE SE DESEMPEÑAN EN SERVICIOS PROFESIONALES Y MANUFACTURA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

CONSTANZA MIRANDA SOTELO

PROFESOR GUÍA:
CRISTIÁN BRAVO ROMÁN

MIEMBROS DE LA COMISIÓN:
PATRICIO VALENZUELA AROS
CARLOS NOTON NORAMBUENA

SANTIAGO DE CHILE
MARZO 2014

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL INDUSTRIAL
POR: CONSTANZA MIRANDA SOTELO
FECHA: 12/03/2014
PROF. GUÍA: SR. CRISTIÁN BRAVO R.

DISEÑO DE MODELOS ECONOMÉTRICOS PARA LA ESTIMACIÓN DE ESTADOS FINANCIEROS DE MICROEMPRESAS QUE SE DESEMPEÑAN EN SERVICIOS PROFESIONALES Y MANUFACTURA

Para BancoEstado Microempresa (BEME) y para la banca en general, las exigencias del mercado y la industria de microempresas en Chile han ido creciendo constantemente llegando a quintuplicarse en los últimos diez años. Es por eso que es muy importante ir mejorando los procesos de créditos y de atención de clientes para funcionar y responder rápidamente a sus demandas y así no perder frente a la competencia.

La Tecnología de Evaluación de Riesgo (TER) es una herramienta que evalúa a los clientes de BEME que necesitan créditos. Actualmente, para la mayoría de los clientes, consiste en visitas en terreno y entrevistas por parte de los ejecutivos con el fin de corroborar la información otorgada por éstos lo cual hace de la TER un procedimiento lento. Con el fin de disminuir este tiempo, se desarrollaron modelos de estimaciones lineales de las variables del estado financiero. De este modo se espera disminuir considerablemente el tiempo que le toma a los ejecutivos realizar un crédito a microempresarios del área Servicios Profesionales y Manufactura.

Se generaron seis modelos de regresión lineal para ventas, costo fijo y margen. Esta memoria presenta las variables que más influyeron sobre el estado de resultado operacional como la formalidad del cliente, la cantidad de empleados de la microempresa, el rubro, si tiene línea de crédito y el tipo de vivienda en la que se encuentra. Estos modelos arrojan una variabilidad porcentual (MAPE) de 2% en promedio para los clientes con historia y de un 10% para los clientes sin historia en el banco. La diferencia entre estos valores está dada en gran parte por la variable que da cuenta de cómo fue el comportamiento de los clientes en el periodo anterior, la cual no está presente en los clientes sin historia.

Finalmente, se proponen alternativas para realizar estos pronósticos utilizando segmentación por rubro de estos, utilizando modelos como probit o logit para modelos que estiman entre el rango (0,1), como el margen, el cual con regresión lineal podría llegar a dar un valores mayores a 1. Otra opción es crear modelos no lineales, como redes neuronales, los cuales pueden captar patrones de comportamiento que no lo hacen los modelos lineales. Se recomienda al banco integrar variables exógenas al modelo, como el PIB o la tasa de desempleo sectorial, de modo que éstos lleguen a ser más robustos.

AGRADECIMIENTOS

Al Centro de Finanzas, de la Universidad de Chile, los cuales me dieron la oportunidad de tomar el tema de memoria que presento a continuación.

A mi comisión, Patricio Valenzuela, y en especial a mi profesor guía, Cristián Bravo, quienes siempre estuvieron dispuestos a ayudarme
en todo lo que fuese necesario.

A BancoEstado Microempresa, por todo el apoyo que tuvieron como contraparte en este trabajo. En especial, a Carolina Venegas quien estaba a cargo de ayudarme con todo lo que fuese necesario.

A mis queridos amigos y compañeros con los que compartí en este periodo, en especial a los que estuvieron conmigo desde mechona. Muchas gracias por los buenos momentos y las alegrías juntos.

A mí adorado equipo de voleibol que me acompaño en toda mi vida universitaria, incluido a Jorge de la Cerda. En la cancha he aprendido cosas igual o incluso más importantes que en las aulas.

Finalmente, a mis padres, mi hermano y toda mi familia.
Gracias por confiar en mí, y ser tan incondicionales

“Intenta no volverte una persona de éxito,
sino volverte una persona de valor.”
Albert Einstein.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

1.1. ANTECEDENTES PREELIMINARES

Las microempresas son unidades económicas de negocios presentes en todo el país y relacionados con todos los tipos de actividad económica. En general, están formados por personas que poseen una baja situación económica por lo que este negocio tiende a ser la principal fuente de ingreso familiar. Son de un tamaño muy pequeño, es decir, poseen pocos trabajadores, pocos activos y un reducido nivel de operación. Según un estudio de la Superintendencia de Bancos e Instituciones Financieras (SBIF) [1] se señala que tienen un bajo nivel de formalización de actividades, donde muchas de estas empresas carecen de iniciación de actividades, patentes municipales o permisos específicos y no tributan impuestos.

La definición más usada de microempresa es la que entrega el Fondo de Solidaridad e Inversión Social (FOSIS) [2]. Esta evalúa tres puntos para calificar una empresa como microempresa:

- Tienen como máximo nueve empleados en total, tanto remunerados como no remunerados, incluyendo al propio microempresario y a sus familiares.
- Su promedio de ventas mensuales es menor a UF 200.
- Tienen activos fijos menores a UF 500.

En Chile, según datos del SII [3] en 2010 la cantidad de microempresas registradas era de 750.555 lo cual representó el 81% del total de compañias. Si se piensa en la cantidad de microempresas que deben existir y que no están inscritas, se cree que la cifra podría alcanzar el millón de empresas fácilmente. En la tabla 1 obtenida del estudio de Morales y Yañez para la SBIF [1] se muestran las características de la clasificación de las empresas en Chile por número de empleados y ventas anuales. En la última columna se puede ver la cantidad de empresas que pertenecen a cada tamaño en el país según el mismo estudio del SII en 2010.

Tamaño	Número de empleados (E)	Ventas anuales UF (V)	Cantidad Empresas
Grandes empresas	E ≥ 250	V ≥ 100.000,1	11.133
Medianas empresas	50 ≤ E ≤ 249	25.000,1 ≤ V≤ 100.000	22.044
Pequeñas empresas	10 ≤ E ≤ 49	2.400,1 ≤ V ≤ 25.000	148.194
Microempresas	E ≤ 9	V ≤ 2.400	750.555

Tabla 1: Tamaño de una empresa según la cantidad de empleados y ventas que posee [1]

En 1995, el Banco del Estado - hoy BancoEstado -, dio un gran paso en la configuración del mapa del microcrédito y las microfinanzas en Chile. Este inició

un programa especializado de microempresas cuyo rol estratégico impactaría en el banco y en el futuro de miles de familias emprendedoras.

Bajo el nombre de BancoEstado Microempresas (BEME) e inspirado en la misión institucional de generar igualdad de oportunidades en el acceso a los servicios financieros para todos los chilenos, el programa era la respuesta para los sectores microempresarios, hasta entonces marginados del sistema financiero. A diferencia de otros bancos, la motivación central para el ingreso de BEME al negocio bancario o de las microfinanzas fue el carácter de banco público de la matriz BancoEstado. Esto último fue lo que hizo que el banco tuviera un especial interés por participar en mercados financieros imperfectos y con impacto social, ya que en esos años existía un acceso al crédito de la microempresa muy inferior a lo que es hoy en la actualidad [4].

Por otro lado, en su creación también incidieron la fuerte competencia en créditos de consumo e hipotecarios que había en los mercados bancarios tradicionales – por lo que era una buena vía para diferenciarse –, la existencia de un mercado desatendido y con grandes posibilidades de crecimiento; y la búsqueda de la diversificación del riesgo. En la década de 1950 no era un argumento muy relevante el atractivo de la rentabilidad ni la evidencia de que los microcréditos fuesen un buen negocio, es más, había muy poca información acerca del negocio de las microfinanzas o créditos para microempresas. Las ONGs eran las que atendían y se hacían cargo de estas entidades.

Durante los últimos años, con el crecimiento sostenido que ha tenido la economía y desarrollo del país, la cantidad de microempresas que ha recurrido al banco a pedir algún tipo de crédito ha aumentado sostenidamente. La ilustración 1, obtenida de la memoria anual 2012 de BEME [5], muestra gráficamente la evolución que ha tenido la cantidad de clientes y el monto total de colocaciones según estudios de BEME.

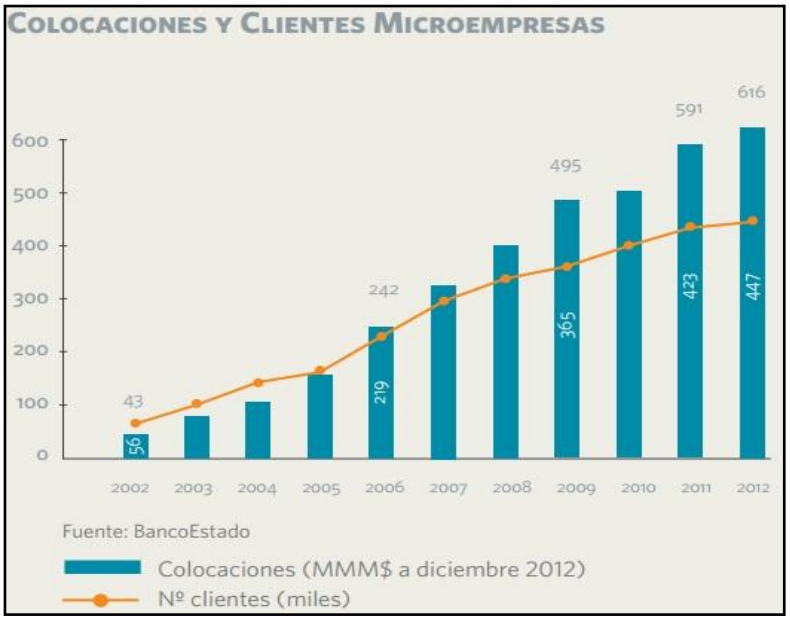


Ilustración 1: Evolución de colocaciones y clientes en Chile según BEME del 2002 al 2012 [5]

Según el informe del Estado de las Microfinanzas en Chile [6], la cartera total de clientes que han accedido a un crédito bancario ha crecido en un 14,7% y en un 8,2% en los años 2011 y 2012 respectivamente. Se estima que la cantidad total de clientes que han accedido a un crédito es de alrededor de 440 mil a lo largo del país. El monto promedio que se otorga en cada una de las ocasiones, es de $1.308.000 aunque se observa un crecimiento importante, en cuanto a la oferta y la demanda, en operaciones con monto promedio por debajo de $1.000.000. Una explicación para esto último, puede deberse a la necesidad de una mayor inclusión financiera entre microempresarios más vulnerables y una cobertura en la oferta a microempresarios que se encuentran consolidados que está llegando al equilibrio de mercado.

En Chile, según el mismo Informe del Estado de las Microfinanzas en Chile [6], las instituciones que otorgan microcréditos son bancos, cooperativas, entidades sin fines de lucro, servicios públicos y sociedades comerciales. La especialización de cada entidad depende del foco de clientes en cuanto a si son urbanas o rurales, sus rangos de ventas o sus niveles de formalidad. El Anexo A se puede encontrar un cuadro con el resumen descriptivo de las 15 instituciones que participan en la Red para el Desarrollo de las Microfinanzas en Chile A.G., la cual agrupa a la mayoría de las instituciones privadas de ahorro y crédito y organismos públicos que ofrecen crédito a microempresas.

BEME presentó una cantidad de clientes atendidos en 2012 de 73.212 y una cantidad de colocaciones en el año de aproximadamente $447 billones, donde el monto de operación promedio fue de $3.716.000. Estos resultados comparados con 2011, presentan una pequeña baja que se cree que se puede estar siendo generada por la competencia que poco a poco ingresa al mercado. Frente a este último escenario, BEME ha optado por mejorar sus servicios, de manera que sean más rápidos y eficientes en cuanto al otorgamiento de créditos y para ello se propone trabajar en una mejora en el área de riesgo, más específicamente en la herramienta TER, la cual se explicará más adelante.

El área de riesgo del banco, se encarga de evaluar el riesgo crediticio de éste por cada préstamo que hace a los microempresarios. Trabaja con modelos de *credit scoring*, entre otros, y se encarga de minimizar la pérdida monetaria que pueda tener el banco.

La Tecnología de Evaluación de Riesgo (TER), es una herramienta que evalúa a los clientes de BEME, que forma parte del proceso de asignación de créditos el cual sigue los siguientes pasos:

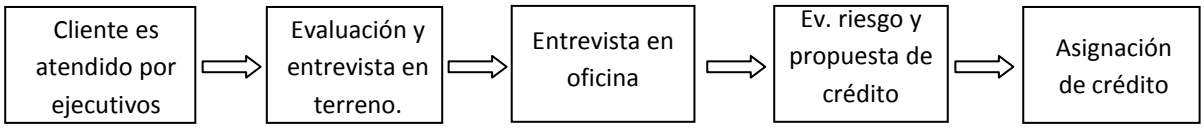


Ilustración 2: Sistema de asignación de créditos bancarios actual.

El proceso actual es el mismo para todos los clientes, ya sean antiguos o nuevos. En primer lugar el cliente es atendido por los ejecutivos bancarios, luego se le da fecha para una entrevista en terreno, donde el ejecutivo comercial visita el lugar de trabajo del microempresario y toma nota acerca de lo que ve y averigua ahí. Finalizado los pasos anteriores, se le pide al cliente que lleve a una segunda reunión con BEME algunos documentos que permitan complementar la información obtenida en terreno como dividendos, cuentas de luz, agua, teléfono, escolaridades, entre otros. Luego de recopilar toda la información obtenida, el ejecutivo de cuentas evalúa al cliente y calcula los estados de resultados de este. Finalmente, el ejecutivo se comunica un par de días después con el cliente y le ofrece un monto final de crédito, el cual se otorga en el momento que el cliente firma el monto en el banco.

El procedimiento actual puede llegar a demorar semanas si es que no se dan las condiciones para que el ejecutivo visite en terreno al cliente. Esto es lo que ha ocurrido bastante en el último tiempo, dado el auge que están teniendo las microempresas. Por esto y por otros motivos que se detallarán en la justificación del proyecto, punto 2.2. de este trabajo, es que se desarrolla TER Express, la cual a través de estimaciones de las variables de estado de resultado, disminuirá y optimizará el proceso actual.

El modelo TER Express que se propone, se puede ver gráficamente a continuación:

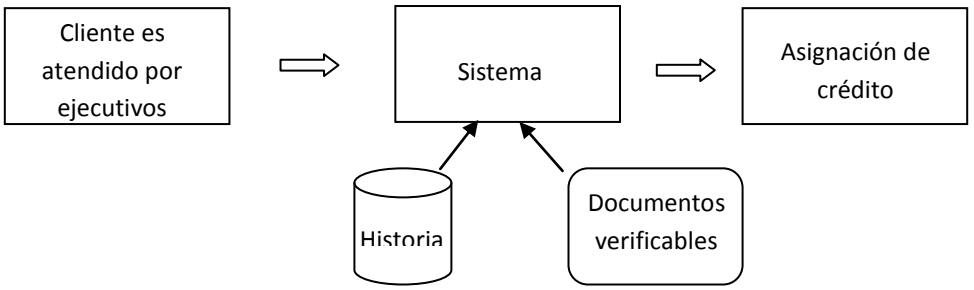


Ilustración 3: Sistema de asignación de créditos bancarios usando TER Express

El diseño de la TER Express consiste en confeccionar modelos de regresión múltiple para estimar las variables más importantes a la hora de calcular el estado de resultados de los clientes, los cuales son la venta, los costos fijos y el margen. Es por ello que para cada uno de estos se ajustarán dos modelos, uno para el caso de que el cliente no presente historia de informes técnicos con el banco (nunca haya pedido crédito o no lo haga en los últimos dos años) y otro pasa el caso que en que el cliente si presente historia de informes técnicos. Para ello el banco otorgará dos bases de datos con los informes técnicos de los clientes que en el periodo de marzo de 2010 a diciembre de 2012 obtuvieron créditos con BEME. Los clientes sin historia corresponden a aquellos que al momento de la evaluación no presentan historia de informes técnicos los dos años anteriores.

Con estas bases de datos se realizan regresiones lineales múltiples para estimar la venta, el costo fijo y el margen. Este modelo consiste relacionar una variable dependiente Y con las variables independientes X, y un término aleatorio ε. Esta relación puede ser expresada como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \varepsilon \qquad (1)$$

En base a estos modelos lineales, se construyen las variables que puedan hacer estimar los estados de resultado operacionales (R.O.) de los clientes de BEME.

$$R.O. = Ventas - Costo\ Fijo - Costo\ Variable \qquad (2)$$

Donde el costo variable se estima a partir del margen[1] dada la siguiente relación:

$$Margen = \frac{Ventas - Costo\ Variable}{Ventas} \qquad (3)$$

El tiempo que tomará realizar estos cálculos permitirá a los ejecutivos comerciales obtener el monto de crédito para sus clientes de una manera mucho más rápida que el método actual. El proceso actualmente dura 1 hora, sin contar los tiempos de traslado a terreno, ni los tiempos en que se conversa con el cliente, solo considerando el traspaso de datos y se espera disminuirlo a 20 minutos, liberando de esta forma 40 minutos de los ejecutivos de cuenta aproximadamente en el procedimiento.

Es importante destacar que este método se implantó el último año para los clientes del sector de Comercio de BEME y en esta memoria se hará el trabajo para poder implementarlo en los sectores de Servicios Profesionales y Manufactura. Se trabajan los modelos de acuerdo al sector industrial al que pertenece el cliente, ya que los comportamientos y los montos asociados son muy distintos en algunos casos.

1.2. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN

El proyecto se traduce en calcular las variables más importantes de los estados de resultado (ventas, costos fijos y margen) mediante modelos de regresión lineal multivariable, usando los informes técnicos de los clientes – que corresponde a la información que el cliente brinda al banco – y la información del sistema del cliente – que corresponde a la información que se encuentra en internet o en otras bases de datos –.

[1] Por lo general el cálculo es al revés, el margen se calcula del costo variable. En este caso es al revés ya que el banco manejas las cifras de los márgenes esperados por segmento de clientes y por lo tanto, le es más fácil comparar.

Para poder hacer los modelos, primero se deberán trabajar las bases de datos, las cuales contaran con 136 variables, realizando algunos filtros y categorizaciones de modo que los modelos no se vean afectados por valores extremos o valores repetidos.

Por otro lado, además de estos modelos lineales, se estudiarán modelos no lineales para ver otras vías de solución a este mismo tema. BEME sugiere y está interesado en los modelos de redes neuronales, pero de todos modos puede que del estudio y análisis que se realice, aparezcan otros modelos como alternativa a solucionar este problema.

Como se planteó anteriormente, este proyecto nace de la necesidad de acelerar la asignación de créditos a los clientes de BEME, ya que el proceso actual toma mucho tiempo y todo ese tiempo se traduce en costos. Los ejecutivos comerciales podrían atender un número mucho mayor de clientes que el que atienden ahora, en donde se pierde mucho tiempo especialmente en las visitas a terreno a los clientes, donde van a corroborar que los datos que entregan los clientes están correctos.

Por otro lado, debido a la gran demanda que están teniendo los créditos para fomentar la pequeña y micro empresa, es que la oferta de créditos a tasas bajas ha ido en aumento por parte de privados. Muchos bancos, entre ellos, BCI, Banco del Desarrollo y Banco de Chile, entre otros, han visto en la microempresa una oportunidad de negocio y por lo tanto han creado áreas internas especializadas en los microcréditos. Esto último ha repercutido en BEME, ya que si bien lideran el mercado de las microfinanzas en Chile, deben ir constantemente mejorando el servicio para poder entrar en la competencia con los bancos privados y atender a todas las microempresas que lo requieran.

1.3. OBJETIVOS

Objetivo General

Maximizar el uso de la información obtenida a través de informes técnicos y el historial sobre datos de clientes, con el fin de proveer seis modelos econométricos de apoyo y orientación a los ejecutivos de BEME, de manera que les permita otorgar créditos a sus clientes más rápida y eficientemente.

1.3.2. Objetivos Específicos

- Conocer el actual sistema de funcionamiento del banco para otorgar créditos y estudiar la herramienta de TER Express, la cual ya está implementada para el sector de Comercio en el banco, para poder aplicarla al sector de Servicios Profesionales y Manufactura
- Generar los seis modelos de regresión multivariable, equivalentes a cada una de las seis bases de datos.

- Verificar que cada uno de los modelos prediga correctamente la variable de respuesta usando algún método de validación.
- Generar un modelo no lineal para alguna de las seis bases de datos y compararlo con el modelo lineal para finalmente concluir cuál es el que predice mejor la variable de respuesta

1.4. METODOLOGÍA

La metodología con la que se desarrolló el trabajo presenta los siguientes pasos:

- **Estudio de modelos TER EXPRESS utilizados para realizar las estimaciones de venta, costo fijo y margen para el área de Comercio.**

Se revisan los documentos pertinentes que tenga el Banco que puedan servir para el desarrollo del tema de memoria. El área de modelamiento predictivo posee un documento llamado Desarrollo de Modelos TER EXPRESS [7] en el cual se muestra el procedimiento que se siguió para realizar estos modelos para el sector de Comercio.

- **Análisis descriptivo de las variables en estudio.**

Usando STATA se estudian las 136 variables que presentan las bases de datos en forma general, de manera que se pueda tener una idea de cómo se comportan los datos en una primera instancia.

- **Limpieza de la base de datos a través de filtros.**

De acuerdo a los resultados que se tienen en el análisis descriptivo de las variables, se realiza una primera limpieza de datos, en donde se eliminan los registros de clientes que presentan información incompleta o que posean valores muy extremos en algunas variables (outliers) dado errores sistemáticos o aleatorios de esta. Por ejemplo, se crea un filtro para la variable “Edad del cliente” en donde se tomarán sólo aquellos clientes que tengan entre 18 y 86 años.

- **Limpieza de variables según la correlación que tengan.**

Luego de tener las 136 variables, se eliminan algunas de estas evaluando la correlación con la variable de respuesta o la auto-correlación entre ellas mismas.

- **Generación de categorizaciones en algunas variables a través de árboles de clasificación, de modo que se simplifiquen los resultados en variables nominales, ordinales y en algunos casos continuas.**

Dado ciertos comportamientos de clientes frente a la variable de respuesta, los resultados de los modelos pueden resultar mejor categorizando las variables y agrupándolas entre aquellos que se comportan de manera similar. Estas categorizaciones pueden ser aplicadas a todo tipo de variables (escalares, nominales y ordinales). Por ejemplo, para la variable "número de niños en el hogar", si se tiene que el número promedio de hijos para los clientes de la base de datos es 3,2; entonces se crearán las categorías agrupadas con respecto a la variable dependiente que corresponda (ventas, costo fijo o margen):

- Sin hijos (0 hijos)
- Pocos hijos (1 o 2 hijos)
- Número promedio hijos (3 o 4 hijos)
- Muchos hijos (más de 5 hijos)

- **Ajuste de los modelos de regresión múltiple**

En este punto se revisa la existencia de colinealidad entre las variables continuas, luego, se estima el modelo con todas las variables disponibles que quedan como "candidatas" para que finalmente a través del algoritmo stepwise, se elijan las variables que finalmente entran al modelo final. Después se revisan algunos supuestos del modelo, como normalidad, homocedasticidad e independencia y se aplican las transformaciones necesarias en caso que haya algún problema con ello.

- **Validación de los modelos ajustados.**

Los modelos se crearán con la base de datos completa y luego se validarán separando la base de datos y dejando una de muestra de entrenamiento y otra de validación. Para esto último, se utilizará el método de validación cruzada aleatoria con k iteraciones. Los parámetros finales del modelo se obtendrán del promedio de los resultados de cada iteración. En el Anexo B se puede ver una imagen que explica este método cualitativamente.

Luego se calculan los indicadores y bondad de ajuste del modelo con la base de datos de validación para comprobar la robustez de los modelos.

- **Estudio de modelo no lineal que pueda adaptarse a la estimación de los cálculos de venta, costos fijos y margen para el problema.**

De acuerdo a los estudios que se encuentren relacionados con el cálculo de estimaciones de crédito a microempresarios, se elabora al final de este trabajo una posible solución no lineal de este mismo problema. Luego se desarrolla este método en una de las bases de datos y finalmente se compararan los resultados con el modelo lineal.

1.5. ALCANCES

Tal como se explicita en el objetivo general, el diseño de los modelos apunta solo a las microempresas que pertenecen a los rubros de Servicios Profesionales y Manufactura. Los otros rubros se dejan afuera ya que según estudios previos del banco, cada rubro suele comportarse de manera distinta y para que los modelos sean lo más exactos posibles, es que se prefiere trabajar por rubros separados y juntar solo los que muestren un comportamiento similar como es el caso de Servicios Profesionales y Manufactura.

Se realizan los modelos de regresión lineal multivariable para costos fijos, ventas y margen en clientes con y sin historia, lo que equivale a realizar 6 modelos. Luego que los modelos estén listos, se podrán calcular estimaciones de los estados de resultados de los clientes del banco.

Por otro lado, se incluye en el estudio la exploración de un modelo no lineal, el cual probablemente será de redes neuronales, de modo que a futuro se pueda dar el puntapié para que en otro estudio se profundice el método para predecir los estados de resultados para microempresas.

2. MARCO CONCEPTUAL

Para comprender y abordar de mejor forma el trabajo es necesario desarrollar ciertos conceptos y proposiciones que se plantean en él, proporcionando una base teórica que lo avale. Se comenzará con unos conceptos básicos de microempresas y del sistema de TER Express y luego se definirán algunos conceptos teóricos importantes para comprender como se desarrolla este trabajo.

2.1. MICROEMPRESAS Y TER EXPRESS

A continuación, se muestran las definiciones de algunos conceptos importantes de entender con relación a las Microempresas y a cómo trabaja BEME con los datos de estas con sus modelos de TER Express.

2.1.1. Tipos de Empresas

Se trabaja con todos los tipos de microempresas que existen en el país. Una clasificación adecuada para evaluar si una empresa es Formal, Semiformal o Informal es la siguiente:

- Empresas Formales: Son aquellas que mantienen registros vigentes en el Sistema de Impuestos Internos (SII) y por lo tanto pagan impuestos. Además poseen todas las patentes, permisos y autorizaciones legales correspondientes. Aparecen de manera segura en los registros de la SBIF.

- Empresas Semiformales: Son las no tienen iniciación de actividades en el SII, es decir, no pagan impuestos; pero si tienen patentes municipales y aparecen registradas con su marca. Estas pueden aparecer o no en la SBIF dependiendo si han pedido crédito antes.

- Empresas Informales: Son aquellas que no tienen registro alguno ni en el SII ni en la municipalidad y por lo tanto no poseen la documentación necesaria para la formalidad y solo poseen cuadernos de registros. Aparecen en los datos del SBIF sólo si han logrado conseguir un crédito en algún banco.

De todas las microempresas que hay en el país, más del 50% de estas corresponden a empresas informales y que se mantienen hasta el día de hoy con fuentes de financiamiento externas mayoritariamente.

2.1.2. Cálculo del Resultado Operacional de la TER Express

En el modelo TER Express se realiza una estimación de las variables dependientes ventas, costo fijo y margen mediante la evaluación express que

hagan los ejecutivos comerciales a los microempresarios y según los modelos de regresión lineal múltiple que se obtengan de este trabajo de memoria para las categorías de Servicios Profesionales y Manufactura. Calculando estas tres variables, es que finalmente se obtiene el Resultado Operacional (RO) de cada uno de los microempresarios.

Luego que se obtiene el RO mediante los dos caminos, se elige el de menor valor y dado ese valor, sumándole tres variables anexas: Otros Ingresos, Gasto Familiar y Deudas (no incluidas antes en los informes técnicos ni en las regresiones lineales), es que se calcula la capacidad de pago.

Esta capacidad de pago es ajustada a ciertos ponderadores, entre ellos la tasa de interés por ejemplo, para dar origen a la capacidad de pago ajustada sobre la cual se origina la oferta de crédito que se le ofrecerá al microempresario. A continuación se puede apreciar el diagrama de funcionamiento de TER Express:

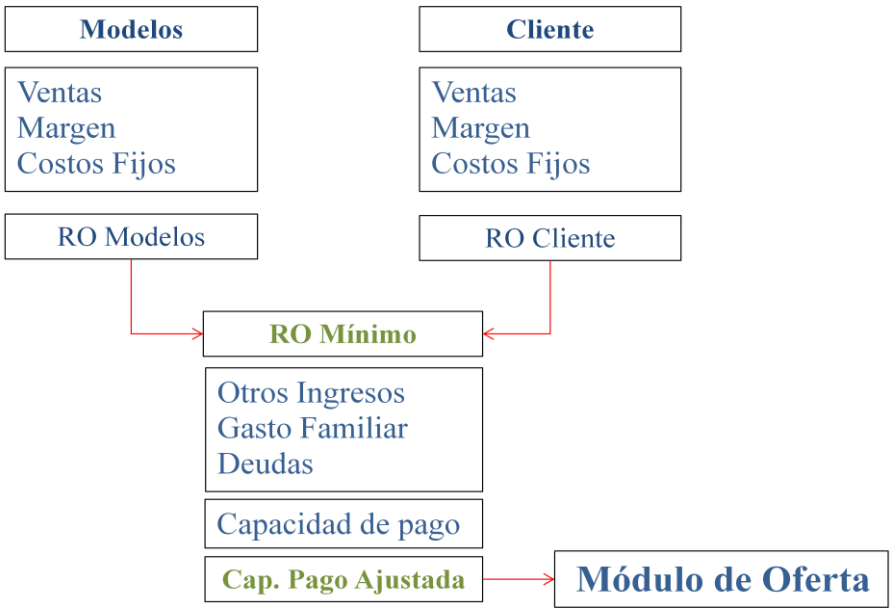


Ilustración 4: Funcionamiento de la TER Express

Es importante decir que si el cliente es totalmente nuevo para el banco, el procedimiento de TER Express no se puede realizar ya que funciona sólo con clientes con historia o clientes sin historia, donde los últimos corresponden a clientes que no piden un crédito hace más de 2 años y no uno que nunca haya pedido crédito antes.

2.2. CLASIFICACIÓN DE VARIABLES

Al poseer dos bases de datos con 8.657 y 18.447 datos de clientes con y sin historia respectivamente y 136 variables, previamente se hará una limpieza de datos previa. Esta consiste en eliminar las variables que no generen valor a la base de datos, como por ejemplo todas aquellas variables que tengan el mismo

valor para todos los clientes. Las variables que presenta la base de datos pueden clasificarse como:

- Cualitativas nominales: Son no numéricas y no presentan un orden. Ejemplo: Estado civil del cliente.

- Cualitativas ordinales: Son no numéricas pero de todas formas si se pueden ordenar. Ejemplo: Perfil del riesgo del cliente puede ser Excelente, Bueno, Regular o Malo.

- Cuantitativa discreta: Con valores aislados (sin intermedios). Ejemplo: Cantidad de hijos.

- Cuantitativa continúa: Estas poseen valores intermedios, como por ejemplo en el caso del valor de las ventas de los clientes.

- Variables dummy cualitativas: Son del tipo de variable dummy que en vez de tener el valor 1 o 0, tiene un valor numérico como Si o No.

2.3. CLASIFICACION DE DATOS

La clasificación estadística de patrones trata el problema de identificar la clase a la que pertenece una observación particular de una población de datos, para la cual su etiqueta de clase es desconocida. Este procedimiento se realiza mediante métodos estadísticos construidos sobre un conjunto de observaciones denominados “base de entrenamiento”, para las cuales sí se conoce su etiqueta de clase [8].

2.3.1. Árboles de Clasificación

Un árbol de decisión es un método de clasificación no paramétrico, cuya estructura es semejante a un diagrama de flujo, en donde cada nodo interno denota un test en cierto atributo, cada rama representa un resultado de ese test, y cada nodo hoja (o terminal) conlleva una etiqueta de clase. La inducción en árboles de decisión es el aprendizaje a través de estas estructuras, a partir de datos de entrenamiento marcados con etiquetas de su clase correspondiente. La popularidad de estos métodos se debe a que la construcción de estos clasificadores no requiere conocimiento del dominio particular desde donde provienen los datos, ni de ajuste de parámetros. Esto los hace particularmente atractivos para análisis exploratorio de datos. Más aún, estos métodos pueden manejar una alta dimensionalidad, y su estructura de representación es intuitiva y fácil de asimilar [9].

Algunos algoritmos populares que implementan el concepto de árboles de clasificación son:

- ID3: usa el criterio de “ganancia de información” para seleccionar atributos.
- CART: usa el criterio del “índice Gini” para seleccionar atributos
- C4.5: usa el criterio de “razón de ganancia” para seleccionar atributos.
- CHAID: usa el criterio “chi-cuadrado” para seleccionar atributos.

En esta memoria se utilizará el algoritmo CHAID, el cual sirve para segmentar los datos. Al igual que otras prácticas de segmentación, las operaciones elementales que ése realiza son: En primer lugar la agrupación de las categorías de las variables predictoras; en segundo lugar la comparación de efectos entre distintas variables, y en tercer lugar la finalización del proceso de segmentación.

2.3.1.1. CHAID

Según el estudio de Araya Alpizar en Segmentación de Mercados [10] y según la memoria de Biron Lattes acerca del Desarrollo y Evaluación de Metodologías para la Aplicación de Regresiones Logísticas [9], el algoritmo CHAID fusiona las categorías de una variable predictora cuando no son significativamente diferentes. Este procedimiento de fusión combinado con el algoritmo de división, asegura que los casos en el mismo segmento sean homogéneos con respecto al criterio de segmentación, mientras que los casos en diferentes segmentos tienden a ser heterogéneos con respecto al criterio de segmentación. Por ejemplo, aunque el número de personas por hogar originalmente podría tener seis categorías; CHAID puede fusionar aquellas cuyos índices de respuesta sean estadísticamente indistinguibles. Por ejemplo, los hogares de dos y tres personas, de cuatro y cinco, se juntaron en una sola categoría. Así, después de la fusión, el tamaño del hogar contiene cuatro categorías lo que sirve para simplificar el modelo final de estimación que se espera lograr. En Anexo D se puede ver un ejemplo gráfico de este algoritmo.

2.3.1.2. Test de Independencia Chi Cuadrado

Este test sirve para comprobar si dos características cualitativas o que estén categorizadas están relacionadas o no, es decir se usa cuando las variables se distribuyen diferente para diversos niveles una de otra. El cálculo proviene de tablas de contingencia en donde se cruzan dos variables con estas características [11].

La hipótesis corresponde a:

H_0: X_1 y X_2 son independientes
H_1: X_1 y X_2 no son independientes

El estadístico chi cuadrado se calcula de la siguiente manera:

$$\chi^2 = \frac{\sum_1^n (Frec\ observada_i - Frec\ esperada_i)^2}{Frec\ esperada_i} \qquad (4)$$

La hipótesis nula se acepta si el valor de chi cuadrado es menor al 0,05 con un nivel de significancia al 95%. Algunos ejemplos donde se utiliza este test, es para las variables de sexo, pertenencia o no de automóvil o maquinas, ciudad donde viven, cantidad de hijos, etc.

Los resultados de este test, permitirán descubrir que variables son independientes entre ellas y por lo tanto permitirán realizar una limpieza de estas.

2.3.2. Redes Neuronales

Una Red Neuronal puede describirse como un modelo de regresión no lineal la cual se inspira en el funcionamiento del sistema nervioso, es decir, es una red con un gran número de unidades simples de proceso, denominas neuronas, que actúan en paralelo y están conectadas mediante vínculos ponderados. En el Anexo C se encuentra una ilustración de un modelo de red neuronal simple.

En las redes neuronales cada conexión se asocia un peso. Durante la fase de aprendizaje, la red aprende mediante el ajuste de los pesos de las conexiones, de manera de predecir correctamente la clase de los individuos de la base de entrenamiento. La gracia de las redes neuronales es que permite identificar sistemas complejos, con no linealidades y dinámica variable, principalmente actuando de dos formas:

- Como clasificador de padrones:

La red es capaz de reconocer eventos como: una perturbación, cambio de propiedades o cambio en el objetivo de control en el sistema y generar una acción correctiva, cambiando los parámetros o modificando la respuesta de un controlador convencional.

- Como modelos de procesos no lineales:

Consiste en usar las redes neuronales como modelo de procesos, e integrarlas a algún esquema de control no lineal tipo MBC (Model Based Control). Los métodos MBC con redes neuronales son los más ampliamente estudiados (Morris, 1994), debido a que es un área bien establecida, con un abundante desarrollo teórico para sistemas lineales, aparte de tener una amplia aceptación en el ámbito industrial.

Existen muchos tipos de redes neuronales, sin embargo, para una posible alternativa de modelamiento, en Anexo C se deja la imagen del tipo de red neuronal Multicapa que es uno de los que posiblemente se podría adaptar al problema que se modela.

2.4. TRANSFORMACION DE VARIABLES

Dado que las variables dependientes de ventas y costos fijos presentan un comportamiento lognormal, se les aplica una transformación para normalizarlas y así poder hacer uso de las regresiones lineales para estimarlas.

Para que se cumpla el supuesto de normalidad de la variable respuesta, la técnica más utilizada para solucionar este problema es realizar una transformación de Box Cox [12].

El criterio para el uso de esta transformación está dado por el cociente entre el valor más grande y más pequeño de Y. Si es de un valor mayor a 10, es posible considerar transformar la variable de respuesta mediante la transformación de Box-Cox. El modelo general tiene la siguiente forma:

$$z(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & si\ \lambda \neq 0 \\ \ln(y) & si\ \lambda = 0 \end{cases} \qquad (5)$$

Para obtener el valor de λ se utiliza el método de máxima verosimilitud. Éste se calcula como sigue para los diferentes valores de λ:

$$U(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \check{y}^{(\lambda-1)}} & si\ \lambda \neq 0 \\ \check{y}\ln(y) & si\ \lambda = 0 \end{cases} \qquad (6)$$

Donde $\check{y}$ es la media geométrica de la muestra pero debido a la magnitud y cantidad de los datos, en este caso se considera la media aritmética. La función de verosimilitud y por lo tanto, la que se maximiza, es:

$$L(\lambda) = -\frac{n}{2}\ln\left(\sum_{i=1}^{n}\left(U_i(\lambda) - \bar{U}(\lambda)\right)^2\right) \qquad (7)$$

Luego se calcula el λ que maximiza el valor de $L(\lambda)$ y dado ese valor, es que se realiza la transformación de la variable dependiente $z(\lambda)$. Los valores más utilizados de λ son los siguientes.

λ	-1	-1/2	0	½	1
Transformación	$z(\lambda) = \frac{1}{y}$	$z(\lambda) = \frac{1}{\sqrt{y}}$	$z(\lambda) = \ln(y)$	$z(\lambda) = \sqrt{y}$	$z(\lambda) = y$

Tabla 2: Transformaciones de Box Cox más usadas

2.5. MODELO DE REGRESION LINEAL

Se espera obtener en este trabajo una regresión lineal dada por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \varepsilon \qquad (1)$$

Donde Y será la variable dependiente o de respuesta, X serán las variables independientes o explicativas, β serán los coeficientes de regresión y, por lo tanto, los parámetros a estimar y ε corresponderá a la perturbación aleatoria que recoja todos los factores que no sean controlables y observables. Esta perturbación aparece por varias razones. La primera razón es que no se puede esperar captar toda la influencia de una variable económica en un modelo, por muy elaborado que este sea. Además existen muchos otros factores que contribuyen a la aparición de dicha perturbación y el más importante corresponde a los errores de medida. Es fácil hablar teóricamente sobre relaciones entre variables definidas con precisión; pero otra cosa es obtener medidas precisas de estas variables. [13]

El problema será buscar los valores determinados para los parámetros desconocidos de todos los β de modo que la ecuación quede completamente especificada. Una de las alternativas para obtener los valores de este parámetro es con mínimos cuadrados ordinarios.

La estimación por mínimos cuadrados ordinarios plantea utilizar como estimación de los parámetros, la combinación de $\beta_1, \beta_2, \ldots, \beta_k$ que minimice los errores que el modelo cometerá. Este vendrá dado por:

$$\varepsilon_i = Y_i - \hat{Y}_i \qquad (8)$$

Donde Y_i es el valor real de la variable dependiente e $\widehat{Y}_i$ es su valor estimado. De acuerdo a esto se tendrá que el β tendrá el valor correspondiente a:

$$E\left(\hat{\beta}_{MCO}\right) = \beta \rightarrow \min \sum_{i=1}^{n} {\varepsilon_i}^2 \qquad (9)$$

Además es importante recordar que este método de mínimos cuadrados es óptimo cuando se cumplen los supuestos del modelo de regresión lineal los cuales se explicarán a continuación.

2.5.1. Linealidad

Dado que el modelo de regresión lineal múltiple se escribe de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \varepsilon \qquad (1)$$

Se tiene que Y es un conjunto de sumas de X multiplicadas por coeficientes de regresión (Betas). Es por eso que la relación de la variable de respuesta Y con respecto a cada uno de sus regresores, debe ser lineal. Para comprobar que esto se cumpla, es que se realiza un análisis scatter plot entre cada uno de los regresores y la variable de respuesta.

El incumplimiento del supuesto de linealidad se denomina error de especificación. Este se puede dar por la no linealidad de la variable de respuesta con respecto a las variables regresoras, por la omisión de variables independientes importantes o la inclusión de variables independientes irrelevantes, entre otros.

2.5.2. Independencia de los errores

La independencia de los errores tiene que ver con la incorrelación de estos, lo que es conocido genéricamente como no correlación entre las perturbaciones o errores. Esto consiste es comprobar que:

$$Cov\left[\varepsilon_i, \varepsilon_j \middle| X\right] = 0 \qquad \forall i \neq j \qquad (10)$$

La falta de independencia, se produce fundamentalmente cuando se trabaja con variables aleatorias que se observan a lo largo del tiempo, esto es, cuando se trabaja con series temporales. Como los modelos con los que se trabajarán en esta memoria corresponden a regresiones lineales que no dependen del tiempo, entonces se asume la incorrelación de estos.

2.5.3. Homocedasticidad en los residuos

El supuesto de homocedasticidad tiene relación con las varianzas de los errores del modelo. Se supone que:

$$Var[\varepsilon_i|X] = \sigma^2 \quad \forall\, i = 1, \dots, n \qquad (11)$$

La homocedasticidad implica que la variación de los residuos sea uniforme en todo el rango de valores de los pronósticos. Hay diversas causas que pueden afectar la homocedasticidad de los residuos. Entre ellas se encuentras: que no se haya incorporado en el modelo alguna variable de importancia, que existan muchos valores extremos o atípicos (outliers) o que las unidades de información, se encuentren particionadas en grupos heterogéneos.

2.5.4. Normalidad de los residuos

Se considera que el valor esperado de un error o perturbación aleatoria debe ser cero para cualquier observación, lo cual se puede expresar de la siguiente forma [13]:

$$E[\varepsilon_i|X] = 0 \quad \forall i \qquad (12)$$

Es decir, la media de cada una de las perturbaciones con respecto a cada variable independiente es cero y si a eso se le agregamos el principio de homocedasticidad, se puede inferir que los errores están normalmente distribuidos con media cero y varianza constante.

Dada esta situación y en base a la cantidad de observaciones que se tiene la ley de los grandes números puede aplicarse y el supuesto de normalidad puede considerarse, al menos aproximadamente, para todos los casos. Se dice entonces que el supuesto de normalidad es innecesario para el modelo de regresión lineal excepto en los casos donde se asume explícitamente alguna distribución alternativa, cuyo caso no es este.

2.5.5. No colinealidad entre variables independientes

Si algunas de las variables independientes X que forman parte de la regresión lineal se pueden expresar como una combinación lineal:

$$a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_k X_k = 0 \qquad (13)$$

Entonces el modelo presenta multicolinealidad. Este problema puede hacer al modelo inestable y afectar la varianza de los estimadores, ya que genera valores más altos en cuanto a la varianza. Este último problema tiene relación con lo difícil de estimar separadamente los efectos marginales o individuales de cada variable explicativa por lo que estos se estiman con poca precisión y por lo tanto el valor estadístico para realizar contrastes de significatividad individual tiende a ser pequeño y se aumenta la posibilidad de rechazar la hipótesis nula, por lo que se tiene a concluir que las variables no son significativas individualmente. Esto ocurre debido a que no se estiman con suficiente precisión los efectos individuales de las variables independientes con alta colinealidad

Existen dos tipos de colinealidad. La primera corresponde a la colinealidad perfecta se da cuando la correlación entre dos o más variables independientes X es muy cercana a 1 y la segunda es la colinealidad parcial corresponde a que la correlación entre dos o más variables independientes sea mayor a 0,7. Para evitar problemas de este tipo, se realizará un análisis de las correlaciones entre todas las variables independientes del modelo y se excluirán las variables redundantes dentro de este.

2.6. ALGORITMO STEPWISE

Un concepto básico importante que se debe conocer, es el algoritmo stepwise con el cual se irán incluyendo variables a los modelos de regresión lineal para poder encontrar las variables que son significantes para el cálculo de ventas,

costos fijos y margen. Este algoritmo según la página del departamento de estadística de Coruña [14] es la mezcla entre:

- Backward Stepwise Regression es el procedimiento que parte del modelo de regresión con todas las variables regresoras y en cada etapa se elimina la variable menos influyente según el contraste individual de la t (o de la F) hasta una cierta regla de parada.

- Fordward Stepwise Regression es el algoritmo que funciona de forma inversa que el anterior, parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa hasta una cierta regla de parada.

- Stepwise comienza como el algoritmo Fordward Stepwise, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer. Este termina cuando ninguna variable entra o sale del modelo.

2.7. INDICADORES DEL MODELO

Luego de general los seis modelos, será de mucha relevancia un análisis de qué tan buenos son para predecir en un próximo periodo las ventas, costo fijo y margen. A continuación se presentan los indicadores que se utilizan.

2.7.1. Bondad de Ajuste y Coeficiente de Determinación

La bondad de ajuste nos permite entender el grado de acoplamiento que existe entre los datos originales y los valores que se obtienen en una regresión. Intentan responder la pregunta de si la variación de los regresores X es un buen predictor de la variación de la variable de respuesta Y.

Hay muchas maneras de medir la bondad de ajuste de un modelo de regresión lineal múltiple. Uno de ellos es usando el coeficiente de determinación (R^2) pero además se puede construir un parámetro que se ajuste a un modelo en especifico según los resultados que arroje.

Considerando $\hat{Y}_i$ como el valor de la variable dependiente dada la regresión lineal múltiple, Y_i el valor real de la variable dependiente e $\bar{Y}$ como el promedio de la variable dependiente Y dada la regresión, se tiene que:

$Y_i - \bar{Y}$	$=$	$\hat{Y}_i - \bar{Y}$	$+$	e_i
Desviación total		Desviación debido a la regresión		Desviación debido al error

Lo cual se puede comprobar con más detalle en la siguiente tabla:

obs	deviation of Y_i	deviation of $\hat{Y}_i = b_0 + b_1 X_{i1} + ... + b_p X_{ip}$	deviation of $e_i = Y_i - \hat{Y}_i$
1	$Y_1 - \bar{Y}$	$\hat{Y}_1 - \bar{Y}$	$e_1 - \bar{e} = e_1$
2	$Y_2 - \bar{Y}$	$\hat{Y}_2 - \bar{Y}$	$e_2 - \bar{e} = e_2$
⋮	⋮	⋮	⋮
n	$Y_n - \bar{Y}$	$\hat{Y}_n - \bar{Y}$	$e_n - \bar{e} = e_n$
Sum of squares	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ Total Sum of squares (SST)	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ Sum of squares due to regression (SSR)	$\sum_{i=1}^{n} e_i^2$ Sum of squares of error/residuals (SSE)

Tabla 3: Errores que se estiman en regresiones lineales.
Fuente: Estudio de la NJIT, New Jersey Institute of Technology

Dado esto, el coeficiente de determinación múltiple se define como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (15)$$

Existen algunos inconvenientes al usar R^2 solamente para medir la bondad de ajuste. Uno de ellos tiene relación con los grados de libertad utilizados en la estimación de los parámetros ya que R^2 nunca decrecerá cuando se añada otra variable a la ecuación de regresión. De esta manera, R^2 en una regresión más amplia no puede ser más pequeño y por lo tanto se puede ver tentado a añadir continuamente variables al modelo. Para evitar esto, es que se presenta el $R^2\ ajustado$ (a los grados de libertad) que se calcula como sigue [13]:

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2)$$

Donde K son los grados de libertad y n la cantidad total de observaciones. El principal beneficio de este nuevo cálculo es que el valor de $\bar{R}^2$ disminuirá (aumentará) cuando se suprima la variable X de la regresión si el estadístico t, de significatividad, asociado a esta variable sea mayor (menor) que 1.

2.7.2. Medidas de Dispersión del Error

- El MAE o Error Absoluto Medio proporciona información acerca del total de los errores en promedio. Esto ya que solo toma el valor absoluto de estos y por lo tanto los errores no se anulan entre sí (positivos y negativos).

$$MAE = \frac{1}{N}\sum_{i=!}^{N} |Y_i - \widehat{Y}_i|$$

- El MSE o Error Cuadrático Medio se define como el promedio del cuadrado de los errores entre el valor real y el pronosticado por un modelo. Este

modelo amplifica los errores más grandes, ya que está elevado al cuadrado. Lo ideal es tener la menor desviación al cuadrado del estimador. por lo tanto, entre modelos, el mejor de ellos es aquel que presenta un menor valor del MSE

–

$$MSE = \frac{1}{N}\sum_{i=!}^{N}\left(Y_i - \widehat{Y_i}\right)^2$$

- El RMSE o la Raíz del Error Cuadrático Medio equivale a la raíz cuadrada de la desviación media o media cuadrática. Es también una medida para comparar modelos entre sí.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=!}^{N}\left(Y_i - \widehat{Y_i}\right)^2}$$

2.7.3. Rendimiento del Pronóstico

- El MAPE o Media Porcentual Absoluta del Error. Este indicador da cuenta del promedio del porcentaje de error, ya que toma el valor absoluto de la diferencia entre el valor real y el pronosticado. Nos dirá qué porcentaje del total de los datos presenta error.

$$MAPE = \frac{1}{N}\sum_{i=!}^{N}\left|\frac{Y_i - \widehat{Y_i}}{Y_i}\right|$$

2.7.4. Relative absolute error (RAE) y Root Relative Squared Error (RRSE)

Las dos medidas que se presentan a continuación, corresponden a medidas de comparación de los modelos con respecto al promedio de todos los datos

- El Error Relativo Absoluto (RAE) mide la diferencia entre el total de los errores o la suma de desviaciones entre los valores reales y el modelo y la division de la sumatoria de la diferencia entre el promedio de los valores reales y sus valores reales.

$$RAE = \frac{\sum|Y_i - \widehat{Y_i}|}{\sum|\bar{Y} - Y_i|}$$

- Para el Error Relativo Cuadrático (RSE), el estimador toma la raíz del cuadrado de los errores, aumentando el valor de los que son más grandes, dividido por la diferencia entre el promedio de los valores reales, como un

modelo simple, y el modelo que se está evaluando, como un modelo más complejo.

$$RRSE = \sqrt{\frac{\sum |Y_i - \widehat{Y}_i|^2}{\sum |\bar{Y} - Y_i|^2}}$$

2.7.5. Correlación

Este valor no dice que tanta relación (directa o inversa) hay entre el pronóstico del modelo y el valor real de la muestra.

$$C_i = \frac{Cov(Y_i, \widehat{Y}_i)}{\sigma_{Y_i}\sigma_{\widehat{Y}_i}}$$

Donde

$$\sigma_{Y_i} = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{N}} \qquad \sigma_{\widehat{Y}_i} = \sqrt{\frac{\sum \left(\widehat{Y}_i - \bar{\widehat{Y}}\right)^2}{N}}$$

Recordemos que Y_i es el valor real de la observación, $\widehat{Y}_i$ es el valor pronosticado por el modelo e $\bar{Y}$ el promedio de los valores reales e $\bar{\widehat{Y}}$ el promedio de los pronósticos. En modelos lineales, el coeficiente de correlación da el mismo valor que el coeficiente de determinación (R squared).

3. PREPARACIÓN DE LA BASE DE DATOS

3.1. DISEÑO BASE ANALÍTICA

Se construye una base de datos analítica con la información contenida en los informes técnicos urbanos que realizan los ejecutivos de BEME. Entre ellos hay información de operaciones cursadas en la banca, datos de la SBIF y la banca, información demográfica, entre otras. Éstas permiten realizar el trabajo de modelamiento predictivo y análisis estadístico para la estimación de las siguientes variables del Segmento Urbano Servicios Profesionales y Manufactura:

- Venta: Se estimará la *Venta* especificada para el mes medio.
- Costo Fijo: Se estimará el *Costo Fijo* especificado para el mes medio.
- Margen: Se estimará el *Margen* especificado para el mes medio. El Margen se calcula a partir de la siguiente fórmula:

$$MARGEN = \frac{\text{VENTA} - COSTO\ VARIABLE}{\text{VENTA}}$$

Donde el *Costo variable* corresponde al mes medio del mes pivote especificado.

Se construirá un Modelo por cada una de las variables dependientes a estimar, y éstos serán explotados en el marco del sistema de evaluación de créditos BEME denominado TER EXPRESS.

3.2. UNIVERSO DE ESTUDIO

Se consideran parte del universo de estudio a los clientes del *segmento Servicios Profesionales y Manufactura* los cuales presenten al menos un informe técnico asociado al mismo rubro en el pivote seleccionado y pasen los filtros seleccionados. La definición anterior garantiza que no aparecen dos Rut repetidos en un mismo pivote.

Debido a la capacidad predictiva de las variables independientes para estimar las variables dependientes (Venta, Costo, Fijo y Margen), se definió a priori una regla que segmente a los clientes en dos categorías, dependiendo de la información de Informes Técnicos del que presenten en el periodo de observación (con el mismo rubro que el Pivote). La regla es la siguiente:

- Categoría SH: Cliente sin historia es aquel que presenta solo un Informe Técnico en el mes pivote respecto a los 24 meses móviles anteriores. La base de datos consta de 18.447 observaciones.

- Categoría CH: Cliente con historia es aquel que presenta el Informe Técnico en el mes pivote y en el periodo de observación presenta al menos un Informe Técnico mismo rubro indicado en el mes pivote. La base de datos consta de 8.684 observaciones.

3.3. HORIZONTE DE TIEMPO CONSIDERADO

Contempla un total de cinco años de información a partir de enero 2008 a diciembre de 2012 los cuales se describe a continuación:

- Periodo de Observación: Considera la información disponible 24 meses anteriores al Informe Técnico del pivote respectivo, además de la historia financiera del cliente disponible al momento de la evaluación según normativa vigente. (enero 2008 a noviembre 2012).

- Pivote: Se define este concepto con el fin de simular el instante de la evaluación del cliente en donde se extrae tanto la información de las variables dependientes o predictoras y ciertas variables independientes del Informe Técnico y de las otras Bases disponibles. En nuestro desarrollo consideraremos 36 meses pivotes. (enero 2010 a diciembre 2012)

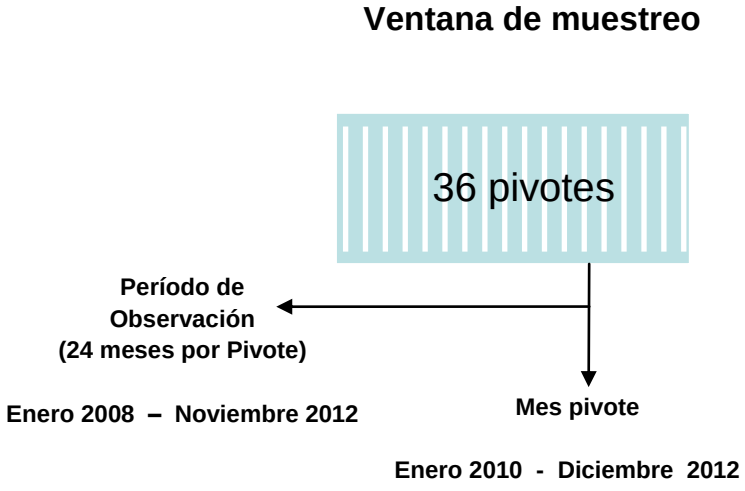


Ilustración 5: Ventana de muestreo de los pivotes de las observaciones
Fuente: Área de riesgo de BEME

3.4. LIMPIEZA DE DATOS

Antes de la construcción de los modelos es necesario realizar una revisión de los datos de modo que se obtengan mejores resultados en estos. Es por ello que se eliminaron algunas observaciones para ciertas variables que estaban incompletas, incorrectas o con valores no pertinentes (outliers) ya sea por errores sistemáticos o aleatorios. Se trabajo con el software de STATA.

A ambas bases de datos de clientes con y sin historia (CH y SH respectivamente), los cuales presentan variables promedio de ventas, promedio de costo fijo y promedio de margen; se les aplican los siguientes filtros:

- Eliminación de clientes con menos de 12 y más de 690 meses en la microempresa, o sin información en ésta.

Esto es equivalente a los clientes que posean una microempresa con menos de un año de antigüedad o más de 57 años de antigüedad. Además se eliminan las observaciones donde esta variable no se especifique. Para la base de datos (bb.dd.) de clientes CH se eliminan 3 observaciones y para la bb.dd. de clientes SH se eliminan 25.

- Eliminación de clientes que tengan menos de 18 y más de 86 años.

Este filtro se realiza para borrar las edades que se pueden haber tipiado mal en el sistema, ya que no se puede pedir crédito si se es menor de edad y por otro lado, no se otorgan microcréditos a personas que sobrepasen los 86 años de edad. Para la bb.dd. de clientes CH se eliminan 126 datos y para la bb.dd. de clientes SH se eliminan 128.

- Eliminación de clientes sin información en valor de vehículo y valor máquina.

Análisis internos del banco estiman que el monto de los vehículos y maquinas con que trabajen los microempresarios, influyen mucho en el poder adquisitivo y en los resultados que tenga un microempresario. Es por esto que se eliminan las observaciones que no tengan valor ahí. Para la bb.dd. de clientes CH se eliminan 10 observaciones y para bb.dd. de clientes SH se eliminan 25.

- Eliminación de clientes que no posean información en formalidad

Se eliminan los registros de clientes que tengan un valor nulo en formalidad, es decir, no tienen definido si son informales, semiformales o formales. Se eliminan 25 observaciones para la bb.dd. CH y 83 observaciones para la bb.dd. de clientes SH.

- Eliminación de clientes que no poseen información en el sistema financiero

Se eliminan las observaciones de clientes que no tienen registro en las bases de la SBIF. Son 8 variables en este campo y se eliminan aquellas que tengan un valor nulo en alguno de los campos. Para la bb.dd. de clientes CH se eliminan 86 observaciones y para bb.dd. de clientes SH se eliminan 758.

Además para clientes SH se eliminan 3.810 datos, los cuales no tenían valor en 33 de las variables de la bb.dd., entre ellas en variables demográficas como sexo, estado civil y edad; y en las variables con la información de la SBIF, relacionadas con deudas directas e indirectas en el sistema (en total 8). Recordemos que la base de datos de clientes sin historia, corresponde a clientes que no piden un crédito hace más de 2 años o clientes nuevo formales. Los clientes informales, tienen que pasar por la evaluación completa para postular a un crédito. Antes de eliminar estos datos, se realiza un análisis de estos con respecto a las pocas variables donde si hay información y se observa un comportamiento similar con la base de datos completa.

La cantidad final de observaciones con que se queda luego de aplicar los primeros filtros, es de 8.434 para clientes CH y 13.618 para clientes SH.

Por otro lado, se agrega otro filtro para ambas bases de datos, el cual consiste en eliminar el 5% de los datos extremos de los datos con respecto a las variables de respuesta ventas, costo fijo y margen. Es aquí donde cada una de las dos bb.dd. se divide en tres, una para cada una de las variables a calcular.

- Eliminación del 5% de los datos extremos por rubro.

La base de datos presenta 10 rubros, cinco para Servicios Profesionales y cinco para Manufactura. De acuerdo a cada rubro es que se eliminó el 5% percentil de las colas según la variable de respuesta (ventas, costo fijo o margen). A continuación se muestra el comportamiento de las colas antes y después de este filtro (el comportamiento para clientes CH o SH es el mismo):

La ilustración 6 muestra el histograma de las variables ventas, costo fijo y margen respectivamente para clientes CH antes y después de realizar los filtros a las observaciones. Hay una clara tendencia en ventas y costo fijo hacia un comportamiento lognormal hacia la cola de valores más altos y para el margen se aprecia un comportamiento bimodal. En Anexo E se puede ver esta ilustración análoga para clientes sin historia, donde las distribuciones son similares y no hay alguna diferencia importante con respecto a los clientes con historia.

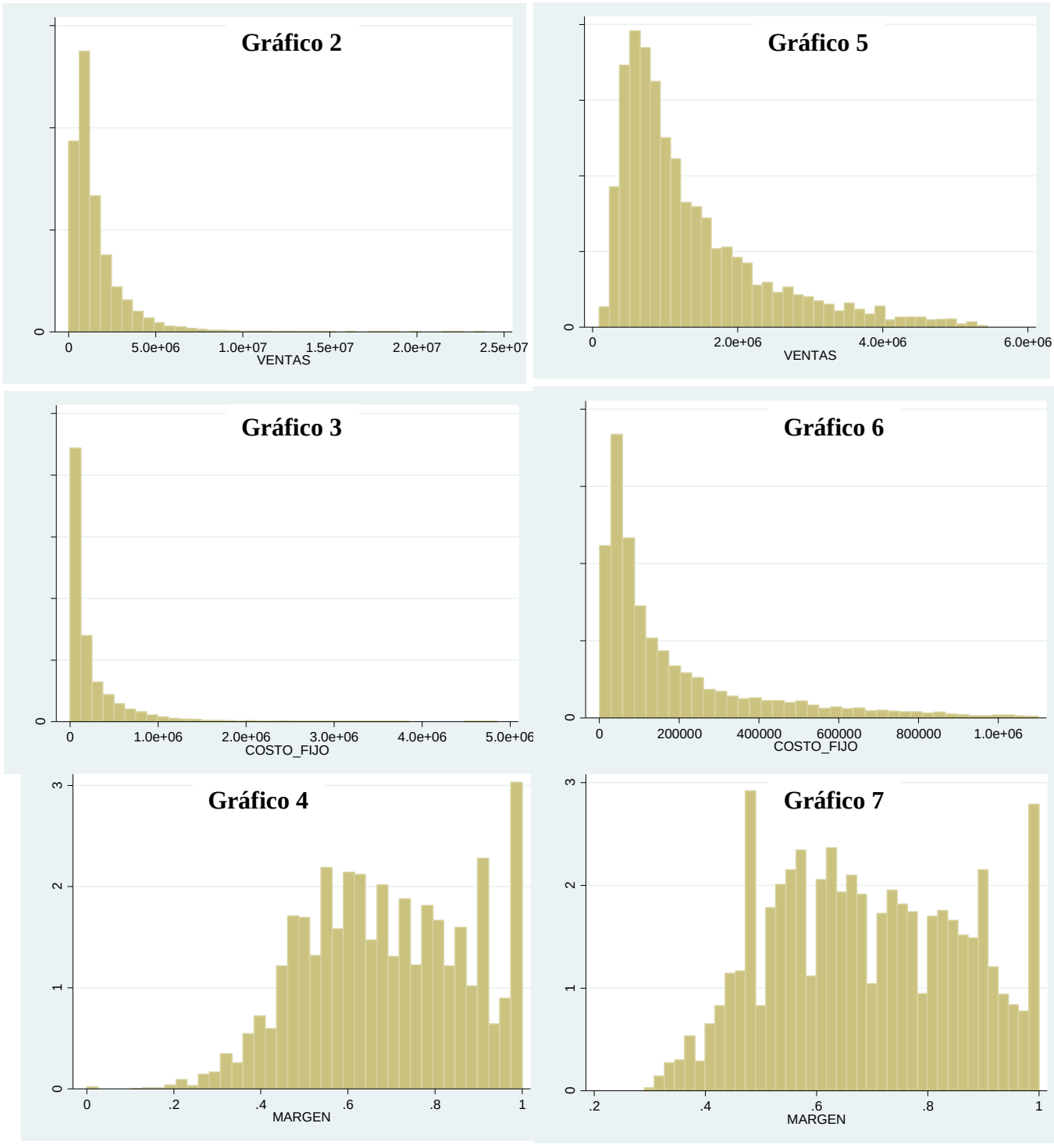


Ilustración 6: Histogramas de las variables antes y después de los filtros
Gráficos 2, 3 y 4 corresponden a histogramas para ventas, costo fijo y margen sin el filtro de percentil. Los gráficos 5, 6 y 7 corresponden a los histogramas luego de aplicar los filtros.
Fuente: Elaboración Propia

Dado que el comportamiento de ventas y costo fijo se puede asimilar a una distribución lognormal es que el filtro de eliminación del 5% de las observaciones se realiza sólo para cola que tiene hacia los datos finales, es decir, la bb.dd. se

queda con el 95% percentil. Se eliminan 426 observaciones en clientes CH y 685 en clientes SH. Como la cola quedaba igual de larga aún, se asigno un tope en cuanto a ventas y costo fijo, eliminándose así todas las observaciones extremas que estuviesen sobre 5,5 millones en ventas y sobre 1 millón en costo fijo.

Para el caso del margen, la bb.dd. se queda con los percentiles 2,5% y 97,5% eliminando colas en los dos extremos. Para clientes CH se eliminan 422 observaciones y para clientes SH, 679.

Para el caso de ventas y costo fijo solo para clientes CH, se agrega además un último filtro. En la limpieza de datos de estas bases, se considera una tasa de ventas, la cual contempla el crecimiento económico mensual de la microempresa, cuyo objetivo es no considerar ventas atípicas relacionadas al segmento. Se define como:

$$\Delta Ventas = \frac{|Ventas - Ventas_{anterior}\ |}{(Antiguedad\ días)/30} \quad \rightarrow \quad Tasa_{ventas} = \frac{\Delta Ventas}{Ventas}$$

Donde:

$|Venta - Venta_{anterior}\ |$: Corresponde a la diferencia absoluta de las ventas.

$(Antiguedad\ días)/30$: Corresponde a la cantidad de meses aproximados entre el IT actual con respecto al IT anterior.

El criterio definido por parte del banco es que si tasa ventas es superior o igual al 7% no se considera en la base de modelamiento. Se borran en total 200 observaciones. Este 7% viene de un estudio previo del banco, donde se observa que los porcentajes de crecimiento de periodo a otro en clientes antiguos para Ventas, no superan el 7% y por lo tanto, las observaciones que lo hagan se estarían comportando como outliers.

Análogamente para el modelo de Costo fijo CH, se considera como filtro de extracción la tasa de Costo Fijo, la cual considera el crecimiento económico mensual de la microempresa, el objetivo de esta tasa es no considerar costos fijos atípicos relacionados al segmento. Se define como:

$$\Delta Costo\_Fijo = \frac{|Costo\_Fijo - Costo\_Fijo_{anterior}\ |}{(Antiguedad\ días)/30} \quad \rightarrow \quad Tasa_{CF} = \frac{\Delta Costo\_Fijo}{Costo\_Fijo}$$

Donde:

$|Costo\_Fijo - Costo\_Fijo_{anterior}\ |$: Corresponde a la diferencia absoluta de los Costos Fijos.

$Antiguedad\ días/30$: Corresponde a la cantidad de meses aproximados entre el IT actual con respecto al IT anterior.

El criterio será que si la tasa costo fijo es superior o igual al 12% no se considera en la base de modelamiento. Se eliminan en total 383 observaciones.

Finalmente, la cantidad de observaciones con las que se trabajará en cada modelo se puede apreciar en la tabla 3. Para las seis bases de datos, se eliminan aproximadamente el 10% de las observaciones iniciales.

Observaciones	Ventas	% obs iniciales	Costo Fijo	% obs iniciales	Margen	% obs iniciales
Con historia	7.791	89,7%	7.563	87,1%	8.012	92,3%
Sin historia	12.915	89.2%	12.907	89.2%	12.939	89,4%

Tabla 4: Cantidad total de observaciones modelamiento

4. CONSTRUCCION MODELOS

4.1. PRIMERA ELIMINACIÓN DE VARIABLES

El primer filtro de variables que se realiza para los modelos, tienen relación con eliminar variables redundantes, es decir, que se repiten dos veces. Por ejemplo, para asignar el rubro de cada uno de los microempresarios, se utilizan 2 variables, una de ellas es el código y otra el nombre del rubro. Es por eso que en todos los casos en que esto ocurra, se deja el código y se elimina la variable string.

El segundo filtro de variables tiene relación con aquellas variables que no aportan información sobre los clientes ya que presentan el mismo valor en todas las observaciones y se eliminan variables como monto cuota, cantidad de cuotas, gastos familiares y otros ingresos ya que con estos datos se ajusta la capacidad de pago del cliente[2]. Para clientes CH y SH se eliminan 6 variables de estos tipos en total.

Por otro lado, para clientes SH se eliminan 41 variables enteras correspondientes a las que se encuentran entre antigüedad días y monto cuota anterior que corresponden a información de ventas, costos fijos y margen anteriores, entre otras, las cuales están en 0 al pertenecer a clientes sin historia y por lo tanto los modelos deben realizarse sin estas variables. La base de datos de clientes CH queda con 110 variables y la de clientes SH con 59

Se procede a calcular la correlación entre todas las variables, es decir los ρ_{xx} y ρ_{xy}. El criterio para eliminar variables será el siguiente: se eliminan todas aquellas variables continuas que tengan un coeficiente de correlación de Pearson entre -0,1 y 0,1 con respecto a la variable de respuesta (ventas, costo fijo o margen) de acuerdo a cada caso. Esto debido a que una correlación tan baja con respecto a la variable de respuesta implica que estadísticamente es muy poco probable que la variable dependiente tenga alguna relación lineal con la variable independiente. Por otro lado, para las variables independientes *X* que tengan una correlación alta entre sí, es decir, un coeficiente de correlación mayor a 0,7 en valor absoluto [15], se dejará solo aquella variable que posea la mayor correlación con respecto a la variable de respuesta *Y*. Esto para evitar problemas de colinealidad entre las variables independientes. La cantidad de variables con las que queda cada modelo es:

	Ventas		**Costo Fijo**		**Margen**	
	Con hist.	Sin hist.	Con hist.	Sin hist.	Con hist.	Sin hist.
Nro variab	45	23	34	21	20	15

Tabla 5: Cantidad de variables finales para usar en cada modelo

[2] Recordemos que la capacidad de pago del cliente es el resultado de la resta entre el Resultado Operacional (R.O.) y estas variables

Para lo que viene a continuación, será importante comprender el significado de algunas variables que para simplificar los resultados, poseen un nombre abreviado.

En Servicios Profesionales y Manufactura existen actualmente 10 rubros asociados a cada uno con su código de abreviación respectivo. Para el primer caso, Servicios Profesionales, se tienen 5 categorías: Belleza y Salud (S1), Educación y Formación (S2), Profesionales y Gestión Empresarial (S3), Alimentación y Amasandería (S4) y Fabricación de Productos Diversos (S5). Por otro lado, para el caso, Manufactura, se tienen las categorías de: Artesanado (M1), Confección Calzado (M2), Mueblista (M3), Reciclaje y Recolector (M4) y Oficios (M5).

Para todas las variables que tengan observaciones en blanco, que son menos del 1% en cada bb.dd, estas se reemplazan por la mediana de cada una. Esto pasa con algunas variables como Puntaje SICA y Puntaje Ambiental. A continuación se muestran los resultados de cada uno de los seis modelos.

4.1 MODELOS PARA VENTAS

4.1.1 Ventas con Historia

Se realiza primero un análisis descriptivo de la variable dependiente, la cual arroja los siguientes resultados:

Mínimo	Moda	Mediana	Media	Max	Desv. Tip.
90.500	1.468.863	1.000.000	1.340.253	5.440.000	996.071

Tabla 6: Estadística descriptiva Ventas con Historia

Luego estos estadísticos se complementan con los valores percentiles, que dan una noción más acabada de la distribución de la variable ventas y confirman su comportamiento log normal. El mayor valor de ventas llega casi a duplicar al valor del 90% percentil. Al ver este comportamiento, es que se decide analizar la mediana por sobre la media al momento de evaluar el modelo ya que esta última está influenciada por los valores que se disparan en la cola.

P5	P10	P25	P50	075	P90	P95
360.000	440.000	625.000	1.000.000	1.739.171	2.800.000	3.540.000

Tabla 7: Percentiles Ventas con Historia ($)

A la variable en estudio, se le aplica la transformación de Box Cox. Para estimar el lambda asociado a la transformación se utiliza Solver de Excel. El resultado arrojado es 0,02 por lo que se decide truncarlo a 0 lo que equivaldría a una transformación logarítmica. La regresión lineal queda del tipo:

$$\ln(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

El histograma luego de aplicar la transformación se puede ver en la ilustración 8.

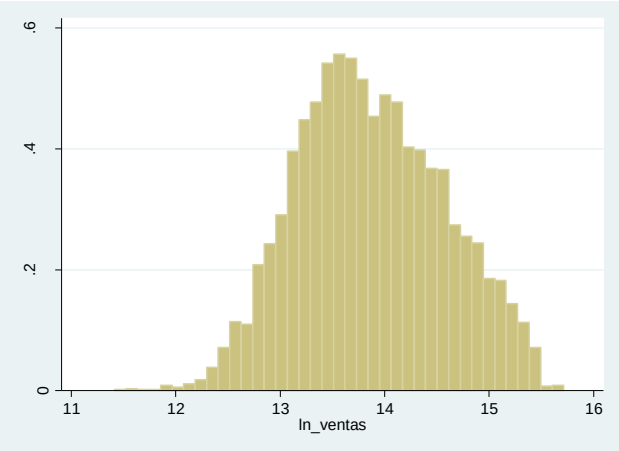


Ilustración 7 : Logaritmo natural de Ventas

Transformada la variable dependiente, se comienza a categorizar con respecto a las variables cualitativas usando el algoritmo CHAID. Para cada variable que se categoriza en n grupos, se crean n-1 variables ficticias (dummies) para ingresar al modelo. En Anexo F se puede ver el detalle de las categorías que quedaron para modelo. Además se comprueba la independencia de estas variables entre sí con el test Chi2 y se elimina una de las variables dependientes.

Para el caso de las variables cuantitativas, existe la variable “venta anterior” que corresponde a la venta que obtuvo el cliente en el periodo anterior al que se trabaja. Esta variable es la que presenta una más alta correlación con la variable dependiente y posee un comportamiento similar (log normal). Es por eso que esta también se transforma a logaritmo natural para ingresarla al modelo junto con la variable de monto de línea de crédito disponible y el monto del vehículo. Los estadísticos descriptivos de estas variables cuantitativas se aprecian en la tabla 8.

Variable	Mínimo	Mediana	Media	Máximo	Desv tip
Línea de crédito	0	0	367,1	12.023	905,5
Deuda consumo comercial	0	1354	2864,3	54.440	4116.0
Puntaje SICA	218,4	906,3	875,2	994	102
Venta anterior	90.500	944.000	1.260.644	9.000.000	966.562,5
Puntaje Ambiental	73,7	778,3	746,3	956,3	126,5
Promedio de deudas	0	55.000	11.6867,6	2.931.189	170.814,7
Monto vehículo	0	0	1.608.718	70.000.000	3.588.697
Monto maquina	0	1.000.000	2.372.863	90.000.000	4.458.545
Margen anterior	0,2	0,7	0,702	1	0,183

Tabla 8: Estadísticos descriptivos variables cuantitativas

Luego de tener todas las variables listas, se ajusta el modelo de regresión lineal. Primero se normalizan todas las variables usando la transformada Z que consiste en restarle el promedio y dividir por la desviación estándar cada una de las observaciones en cada variable como se muestra a continuación:

$$Z_{xi} = \frac{X_i - \bar{X}}{\sigma_x}$$

Luego se divide la base de datos en una parte de entrenamiento y otra de validación con un 70% y un 30% para cada parte. Se utiliza la validación cruzada aleatoria con 10 iteraciones para asegurar que el modelo quede robusto. El algoritmo utilizado para la selección de variables es el Stepwise. Los resultados de la regresión lineal son los siguientes:

Atributo	Beta	Std Error	T-stud	P-valor
Monto máquina	0.0104	0.0024	4.23	2.42E-5
Niños	0.0122	0.0025	4.90	9.77-7
Adultos	-0.0101	0.0024	-3.99	6.75E-5
Deuda consumo comercial	0.0193	0.0033	5.80	6.63E-9
Margen anterior	-0.0197	0.0024	-7.79	7.43E-15
Promedio de deudas	-0.0068	0.00302	-2.17	0.0339
Si tiene línea de crédito	0.0101	0.00264	3.82	1.39E-4
Vivienda: Arriendo o propia con deuda	0.01084	0.0024	4.45	8.7E-6
Formalidad: Informal	-0.01852	0.0027	-6.95	3.89E-12
Región oficina G2	0.00751	0.0023	3.33	9.25E-4
Categoría rubro G1	-0.0073	0.0023	-2.94	0.00346
Ln venta anterior	0.5992	0.00393	152.57	0.0
Ln monto vehículo	0.0122	0.0025	4.85	1.24E-6
Línea de crédito disponible	0.0132	0.00273	4.83	1.37E-6
Empleados: 2	0.01975	0.00245	8.07	7.77E-16
Empleados: 3 ó más	0.0431	0.00283	15.23	0.0
Puntaje SICA	0.016	0.0026	4.436	9.57E-6
Puntaje ambiental	-0.0084	0.0024	-3.423	6.61E-4
CONSTANTE	13.87	0.0022	6287.07	0.0

Indicador	MAE	MAPE	MSE	RMSE	RAE	RRSE	R^2
Valor Modelo	0.128	0.93%	0.038	0.195	0.223	0.279	0.96

Tabla 9: Regresión lineal e Indicadores del modelo de Ventas con historia

Se puede ver que claramente la variable que más influye en el modelo de ventas con historia es la venta que tuvo el cliente en el periodo anterior (en rojo). Por otro lado, los signos de los coeficientes de las variables que entran en el modelo son coherentes con lo esperado. El grupo 1 de categoría rubro,

conformado por Artesanado, Confeccionista de Calzado y Reciclaje-Recolectores, era el que tenía menor promedio de ventas, por lo tanto el coeficiente que aporta se resta a la constante, la cual incluye el aporte de la categoría 2 libre (otros 7 rubros). También se puede apreciar que los clientes Informales restan a la constante, lo que es muy lógico ya que los Formales y Semiformales tenían un promedio de ventas mucho mayor. Por otro lado, es correcto suponer que si el monto de la línea de crédito disponible es mayor, es debido a que las ventas son mayores o viceversa, pero hay una relación positiva.

Otra de las variables más influyentes en el modelo tiene relación con la cantidad de empleados que posee la microempresa del cliente. Ya con dos empleados las ventas son mayores que con uno y si tiene tres o más, el coeficiente es mucho mayor aún por lo tanto, hay una proporción directa entre empleados y ventas.

Para validar estos modelos se utilizan los indicadores vistos en el marco conceptual. El coeficiente de determinación da de un 96%, lo que quiere decir que el modelo explica un 96% de la variabilidad. Por otro lado el valor del MAPE y RRSE que son indicadores porcentuales, dan cuenta de la poca variabilidad porcentual que tiene el modelo. Finalmente El MAE con el MSE da cuenta del bajo valor de diferencia que hay en promedio entre el modelo y los valores reales y finalmente el RMSE es equivalente a la desviación estándar del error entre el pronóstico y valor real de las ventas.

Adicionalmente a estos indicadores, se calcula la sobre y subestimación del modelo con un nivel de significancia de un 5% para la base de datos de muestra y para la de validación. En la tabla 10 se pueden ver los resultados. Si bien el modelo tiene unos excelentes indicadores, al aplicar exponencial a la variable de respuesta, que recordemos que está con logaritmo natural, los valores se disparan y no son tan buenos como se esperaba.

	Entrenamiento		**Validación**	
Estimación v/s Ventas	**N**	**%**	**N**	**%**
Sobre 5% estimación	2366	43,8%	1012	42,2%
Entre 95% y 105%	1665	30,9%	744	31,1%
Bajo 5% estimación	1365	25,3%	639	26,7%
Total	5396	100%	2395	100%

Tabla 10: Indicador MAPE para sobre y subestimación

La tabla 10 muestra el MAPE y calcula los casos en que la diferencia entre el valor estimado y el real es menor al 5%. Esto último ocurre en el 31% de los casos aproximadamente.

4.1.2 Ventas sin Historia

Al igual que para el modelo de ventas con historia, en este caso se realiza un análisis descriptivo de las variables

Mínimo	Moda	Mediana	Media	Max	Desv. Tip.
0	1.100.000	1.050.000	1.429.734	5.500.000	1.095.600

Tabla 11 Estadística descriptiva Ventas sin Historia

Se observa que la Moda y la Mediana están bajo el valor de la Media. Esto a diferencia del modelo de ventas con historia, da cuenta que el comportamiento de clientes sin historia es más pronunciado que el anterior, es decir, es una log normal con una cola mucho más larga. En Anexo E se puede ver el histograma para corroborar esta información y en la tabla 12 se ve que los valores en percentiles también son mayores a los con historia del punto anterior en aproximadamente un 10%.

P5	P10	P25	P50	075	P90	P95
345.000	430.000	640.000	1.050.000	1.890.000	3.026.240	3.897.832

Tabla 12: Percentiles Ventas sin Historia

A la variable en estudio, se le aplica la transformación de Box Cox. Para estimar el λ asociado a la transformación se utiliza Solver de Excel. El resultado arrojado es 0,4 lo que la regresión queda de la siguiente forma:

$$\frac{y^{\lambda}-1}{\lambda} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

El histograma luego de aplicar la transformación se puede ver en la ilustración 8.

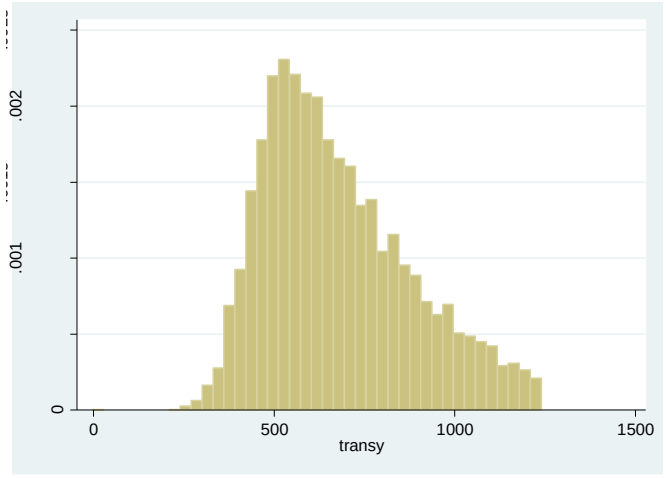


Ilustración 8 Transformación Box Cox de Ventas

Luego de esto, se comienza a categorizar con respecto a las variables cualitativas usando el algoritmo CHAID. En Anexo F 2) se puede ver el detalle de las categorías que quedaron para este modelo. También se comprueba la independencia de estas variables entre sí con el test Chi2 para ver si hay dependencia entre variables y tener que eliminar la que tenga menos correlación con la variable de respuesta. En este caso no hay variables dependientes.

En Anexo G, se pueden apreciar los estadísticos descriptivos de las variables cuantitativas del modelo, que son casi las mismas que el modelo anterior. En el caso de ventas de clientes sin historia, no hay ninguna variable del periodo anterior (como venta anterior para clientes con historia) que pueda explicar el comportamiento de la variable de respuesta. La variable que tiene una correlación más alta es el monto del vehículo que tiene beta positivo y pertenencia de vehículo (1, si tiene y 0, si no) la cual tiene beta negativo. Es extraño tener dos variables relacionadas con algo similar, tenencia y valor de vehículo, con betas tan opuestos. Al contrastar ambas variables por si solas con el modelo de respuesta nos da lo siguiente:

$$\check{y} = 13{,}654 + 0{,}611 * X_{pertenencia\ \ vehiculo}$$

$$\check{y} = 13{,}78 + 6{,}46E^{-8} * X_{valor\ \ vehiculo}$$

Si además la correlación entre ambas variables independientes es 0,546; entonces es de suponer que hay colinealidad en el modelo y por eso da esos valore en distintas direcciones. Se elimina la que tiene menor correlación con la variable de respuesta y, que es monto vehículo, se normalizan las variables (transformada Z) y se vuelve a correr el modelo.

En la tabla 13 se pueden ver los coeficientes del modelo estimado. Se realiza nuevamente la validación cruzada aleatoria con 10 iteraciones para obtenerlo, por lo tanto, los coeficientes son el resultado del promedio de todas estas iteraciones. Nuevamente dan coherente los coeficientes y el valor de permanencia de vehículo disminuye considerablemente con respecto a cómo daba cuando se encontraba con el monto del vehículo. El modelo posee básicamente las mismas variables que el con historia, sin embargo al no tener una variable tan fuerte como venta anterior, hace que entren otras variables como DDVI (Deuda Directa Vigente según la SBIF) donde queda que a mayor deuda vigente, menor pronóstico para las ventas. Una diferencia con el modelo anterior es que los coeficientes de los informes de puntaje ambiental y puntaje SICA son ambos negativos (en el modelo con historia, puntaje SICA dio un valor positivo).

Nuevamente entra al modelo las variables relacionadas con la cantidad de empleados como las más importantes, se ve una proporción directa entre el costo fijo estimado y la cantidad de empleados. Otra de las variables más influyentes es la deuda máxima la cual es coherente con que mientras mayor sea el valor de esta deuda, mayor sean los costos de la microempresa.

El coeficiente de determinación es de un 58,1%, lo que quiere decir que el modelo explica ese porcentaje de la variabilidad. Es mucho más bajo que el modelo anterior debido a que como ya se ha dicho, no hay ninguna variable con una correlación alta (sobre 0,5) con la variable de respuesta. Los otros indicadores muestran que el error es mayor en este modelo, cosa que ya se esperaba, pero tampoco es tan malo, por ejemplo el MAE es aproximadamente el doble y el RRSE, da cuenta que el modelo es muy preferible por sobre estimar con el promedio. Para más detalle observar los indicadores de la tabla 13.

Atributo	Beta	Std error	T-stud	P-valor
Monto máquina	0.0083	0.0046	17.077	0.0
Si tiene vehículo	-0.111	0.0606	-14.19	0.0
Niños	0.055	0.0044	11.99	0.0
Adultos	-0.045	0.00442	-10.00	0.0
Deuda máxima	0.264	0.0377	6.106	7.82E-13
Deuda Directa Vigente (DDVI)	-0.208	0.0374	-4.912	1.134E-8
Puntaje SICA	-0.021	0.0044	-5.127	7.81E-7
Si tiene línea de crédito	0.074	0.0048	13.82	0.0
Ln línea de crédito disponible	0.052	0.0049	9.143	0.0
Cat rubro G1	-0.011	0.0055	-2.6552	0.0087
Cat rubro G3	-0.056	0.0055	-10.452	0.0
Vivienda G2	0.0434	0.0044	9.841	0.0
Empleados: 2	0.152	0.0047	32.789	0.0
Empleados: 3 ó más	0.299	0.005	58.632	0.0
Formalidad: Informal	-0.146	0.005	-28.671	0.0
Puntaje ambiental	-0.009	0.0045	-2.6526	0.037
CONSTANTE	13.91	0.0043	3241.11	0.0

Indicador	MAE	MAPE	MSE	RMSE	RAE	RRSE	R^2
Valor Modelo	0.238	2.73%	0.227	0.476	0.624	0.654	0.581

Tabla 13: Regresión lineal e Indicadores de modelo de Ventas sin historia

Se calcula además la sobre y subestimación del modelo con un nivel de significancia de un 5% para la base de datos de muestra y para la de validación (70% y 30% respectivamente). En la tabla 14 se pueden ver los resultados del MAPE dentro de un 5% de variación con respecto a Y. Es importante considerar que en este caso, como estamos hablando de estimar ventas, el banco prefiere subestimar más que sobrestimar. Esto debido a que mientras menores sean las ventas, la capacidad de pago es menor y viceversa, por lo que en caso de sobre estimar se podría dar en algunos casos más crédito que lo que puedan pagar.

	Entrenamiento		Validación	
Estimación v/s Ventas	**N**	**%**	**N**	**%**
Sobre 5% estimación	4184	46,4%	1833	46,8%
Entre 95% y 105%	770	8,6%	313	8%
Bajo 5% estimación	4047	45%	1768	45,2%
Total	9001	100%	3914	100%

Tabla 14: Indicador MAPE para sobre y subestimación

La tabla 14 muestra el MAPE y calcula los casos en que la diferencia entre el valor estimado y el real es menor al 5%. Esto último ocurre en el 8,4% aproximadamente de los casos, si promediamos ponderadamente los resultados de la base de muestra de entrenamiento y validación.

Se intentó mejorar el valor del modelo, ingresando una variable ficticia de venta anterior, con respecto a los valores de ventas con historia. Se tomó el promedio de esta variable por empleado y formalidad y se añadió como una variable extra a la base de datos de clientes sin historia. La tabla 15 muestra la variable que se ingresó.

ln vta anterior	**1 empleado**	**2 empleados**	**3 o más empleados**
Formal	13.76584	14.12035	14.6109
Semiformal	13.51221	13.89677	14.38847
Informal	13.21778	13.62713	14.04316

Tabla 15: Variable ficticia de prueba para modelo Ventas sin historia

El resultado no fue significante, ya que si bien, esta variable era la que mas explica este nuevo modelo e ingresa al modelo al comienzo, toma el lugar de Empleados y Formalidad (dejándolas fuera del modelo) pero los indicadores son los mismos y no mejoran en casi nada. Se probó de manera similar para los modelos de costo fijo y margen sin historia y ocurrió algo similar.

4.2. MODELOS PARA COSTO FIJO

4.2.1 Costo Fijo con Historia

Se realiza primero un análisis descriptivo de la variable dependiente, la cual arroja los siguientes resultados:

Mínimo	**Moda**	**Mediana**	**Media**	**Max**	**Desv. Tip.**
1.000	50.011	95.000	179.022,2	1.000.000	201.230,6

Tabla 16: Estadística descriptiva Costo Fijo con Historia ($)

Luego estos estadísticos se complementan con los valores percentiles, que se encuentran en la tabla 19, y dan una noción más acabada de la distribución de la variable costos fijos. Se ve como los valores aumentan exponencialmente a medida que crece y como la moda y mediana están muy por debajo de la media, lo que significa que la media está muy afectada por la larga cola de distribución pero no representa necesariamente a la mayoría de las observaciones.

P5	P10	P25	P50	075	P90	P95
17.000	25.000	45.000	95.000	235.000	480.811	648.000

Tabla 17: Percentiles Costo Fijo con Historia ($)

La estimación del λ para la transformación de Box Cox arroja un λ igual a 0, por lo tanto la regresión lineal queda de la siguiente manera:

$$\ln(y_i + 1) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

La ilustración 9 muestra el histograma de los costos fijos, luego que se aplica la transformación de logaritmo natural. El corte que aparece en la cola derecha es producto de la distribución tan marcada de la variable. En la ilustración 6 se puede ver como es el histograma de costo fijo.

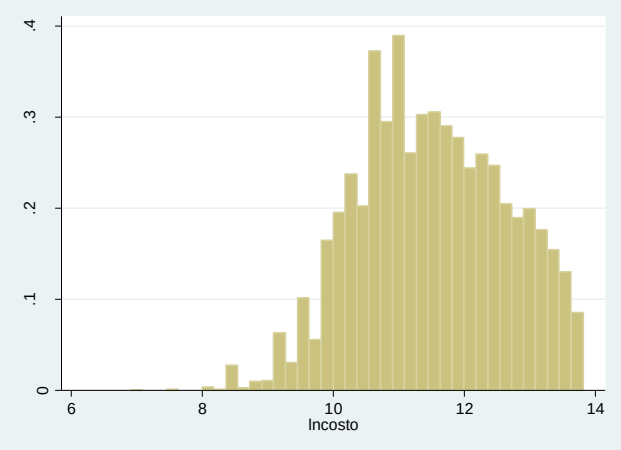


Ilustración 9: Logaritmo natural del Costo Fijo

Luego se categorizan las variables cualitativas y se les aplica el test de independencia. Hay variables que son dependientes pero que no tienen una correlación muy significativa con la variable de respuesta y por lo tanto ninguna entra en el modelo final.

Para el caso de las variables cuantitativas, existe la variable “venta anterior” que corresponde a la venta que obtuvo el cliente en el periodo anterior al que se trabaja. Esta variable es la que presenta una más alta correlación con la variable dependiente y posee un comportamiento similar (log normal). Es por eso que esta

también se transforma a logaritmo natural para ingresarla al modelo junto con la variable de monto de línea de crédito disponible y el monto del vehículo. Los estadísticos descriptivos de estas variables cuantitativas y continuas se aprecian en la tabla 18.

Variable	Mínimo	Mediana	Media	Máximo	Desv tip
Costo fijo anterior	0	80.000	162.303,6	1.820.000	196.493,4
Monto total anterior	101.220	1.049.343	1.729.561	3,13e+07	2.161.921
Puntaje SICA	196.7	906.2	875.037	994	102.3
Venta baja anterior	0	720.000	979.972	1,18e+07	828.004,4
Puntaje Ambiental	73.7	778,3	746,3	956,3	127.7
Periodo anterior	0	61.380	133.432,56	2.851.866	201.666
Monto vehículo	0	0	1.663.198	70.000.000	3.660.625
Monto máquina	0	1.034.000	2.373.860	90.000.000	4.381.142
Plazo op 12 m	0	12	16,8	96	15,3
Plazo op 24 m	0	18	21,1	96	13,4
Gasto fijo prom	70.000	194.500	213.667,3	1.400.000	97.534,1

Tabla 18: Estadísticos descriptivos variables cuantitativas y continuas

Se normalizan las variables, según la transformada Z. La fórmula es la siguiente:

$$Z_{xi} = \frac{X_i - \bar{X}}{\sigma_x}$$

Luego se divide la base de datos en una parte de entrenamiento y otra de validación con un 70% y un 30% para cada parte. Se utiliza la validación cruzada aleatoria con 10 iteraciones para asegurar que el modelo quede robusto. El algoritmo utilizado para la selección de variables es el Stepwise. Los resultados de la regresión lineal se pueden ver en la tabla 19. El modelo queda con más variables continuas que categóricas, a diferencia de los modelos de ventas. Nuevamente se ve que la variable del periodo anterior relacionada con la variable de respuesta, es la que tiene una mayor importancia en el modelo. En este caso es el costo fijo anterior, el cual posee un beta de 0,805. Otra de las variables importantes de este modelo, tiene relación con la cantidad de empleados, donde es de esperar que a medida que hay más empleados, el costo fijo sea mayor (salarios). La variable deudas it anterior, corresponde a la suma de todas las mensualidades medias contraídas por la microempresa en el periodo anterior y por lo tanto si tiene un beta positivo, significa que tiene más deudas y por lo tanto más costos.

Los indicadores estadísticos, que se ven al final de la tabla 19, muestran que el modelo tiene poco error porcentualmente. Si bien el R^2 es mayor que el del modelo de ventas con historia, tener un 83% se considera un muy buen valor. En cuanto al Error Absoluto (MAE) y Error Cuadrático Medio (MSE), se tiene un muy bajo valor de la suma de las desviaciones, por lo tanto los valores pronosticados en logaritmo natural, se parecen mucho a los reales.

Atributo	Beta	Std error	T-stud	P-valor
Monto máquina	0.0293	0.006	4.78	1.83E-6
Monto total anterior	0.025	0.0069	3.632	2.98E-4
Si tiene línea de crédito	0.023	0.0062	3.7	2.253E-4
Ln venta baja anterior	0.055	0.0061	8.96	0.0
Ln costo fijo anterior	0.805	0.0075	107.64	0.0
Ln deudas anteriores	0.016	0.006	-2.17	0.035
Ln monto vehículo	0.034	0.0062	5.55	3.073E-8
Empleados: 2	0.089	0.006	14.96	0.0
Empleados: 3 o más	0.143	0.0067	21.32	0.0
Modulo: G1	-0.0321	0.0055	-5.86	4.87E-9
Cat rubro: G2	-0.056	0.0059	-9.59	0.0
Vivienda Propia con deuda o Arriendo	0.021	0.0058	3.201	0.0014
Formalidad: Formal	0.062	0.0065	9.63	0.0
Puntaje ambiental	-0.017	0.00568	-3.016	0.0028
Puntaje SICA	0.025	0.00619	4.07	4.77E-5
CONSTANTE	11.51	0.0054	2134.11	0.0

Indicador	MAE	MAPE	MSE	RMSE	RAE	RRSE	R^2
Valor modelo	0.299	2.66%	0.22	0.469	0.32	0.416	0.83

Tabla 19: Regresión Lineal e Indicadores modelo Costo Fijo

El valor del MAPE, cuando la variable de respuesta está en logaritmo, es muy bajo y por lo tanto es un buen valor. Sin embargo, al aplicar exponencial y pasar los valores de costo a los originales, el valor del MAPE cambia drásticamente. Esto se puede ver en la tabla 20 donde se analiza la sobre y subestimación. A diferencia del caso de ventas, en costo fijo se prefiere la sobre estimación en vez de la sub estimación. Esto ya que el costo es inversamente proporcional a la capacidad de pago final generada por el modelo de TER Express, por lo tanto si se sub estima, puede generar una capacidad de pago mayor a la que pueda pagar un microempresario.

	Entrenamiento		Validación	
Estimación v/s Ventas	**N**	**%**	**N**	**%**
Sobre 5% estimación	2453	46,3%	1080	47,6%
Entre 95% y 105%	678	12,9%	289	12,7%
Bajo 5% estimación	2163	40,8%	901	39,7%
Total	5295	100%	2270	100%

Tabla 20: Indicador MAPE para sobre y subestimación

El 12,8% de los pronósticos se encuentra con una diferencia de menos del 5% del valor real de costo fijo. Si bien, no es un muy buen valor, se puede ver que un 47% de los pronósticos están sobrestimados, lo cual, hace que al momento de calcular un Resultado Operacional (R.O), es decir, las ventas menos los costos fijos y variables, se tenga un resultado más bajo, lo que es preferible por sobre tener un R.O. alto que podría poner en riesgo al banco frente a clientes riesgosos.

4.2.2 Costo Fijo sin Historia

Al igual que para el modelo de ventas con historia, en este caso se realiza un análisis descriptivo de las variables. En la tabla 24 se muestran los estadísticos descriptivos más relevantes de la variable en observación.

Mínimo	Moda	Mediana	Media	Max	Desv. Tip.
0	50.000	103.844	201.351,6	1.200.000	232465,5

Tabla 21: Estadística descriptiva Costo fijo sin Historia ($)

El costo fijo presenta un comportamiento log normal, al igual que en clientes con historia. Del percentil 95 al valor máximo, se puede ver que casi duplica el último al primero y por lo tanto posee una cola muy larga hacia los valores más altos. Por otro lado, hay costo mínimo 0, lo cual es extraño pero puede darse en Servicios Profesionales, en casos de albañiles o educadores que posean sólo costos variables. Se recuerda que los microempresarios de Servicios Profesionales y Manufactura trabajan realizando oficios por lo tanto puede ser que no tengan algún costo fijo dentro de su flujo de ingresos.

P5	P10	P25	P50	075	P90	P95
15.000	25.000	45.000	103.844	270.000	530.000	730.000

Tabla 22: Percentiles Costo fijo sin Historia ($)

La estimación del λ para la transformación de Box Cox nuevamente arroja un lambda igual a 0, por lo tanto, la regresión lineal queda de la siguiente manera:

$$\ln(y_i + 1) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

Se le suma 1 a la variable dentro del logaritmo natural, para que los valores que tienen cero no se indefinan en la transformación. Esto afecta a la curva normal generada, si se observa la ilustración 10, se puede ver que hacia la izquierda hay valores en 0 que distorsionan un poco esta curva.

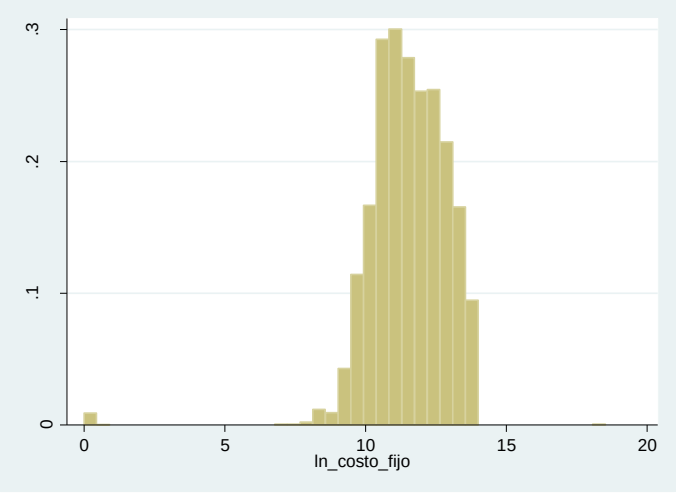


Ilustración 10: Logaritmo natural del Costo Fijo sin Historia

El siguiente paso es la categorización. En Anexo F 4) se pueden ver las categorías de variables cualitativas que se generan con respecto al costo fijo. En este modelo no existe alguna variable relacionada con el periodo anterior y ocurre algo similar al modelo de ventas sin historia, no hay ninguna variable que tenga una correlación alta (sobre 0,7) con costo fijo. En Anexo G se encuentran los estadísticos descriptivos de las variables continuas que son candidatas al modelo, la cantidad de variables continuas que hay son mucho menos que en el modelo con historia, donde habían tres variables continuas que tenían relación con el periodo anterior.

Luego de normalizar y de que se tome el promedio de las 10 iteraciones de validación cruzada aleatoria, se tiene en la tabla 23 el modelo con los coeficientes de estimación de costo fijo. Al desarrollarlo, se detectó nuevamente un problema de colinealidad entre dos variables que entraban al modelo: Módulo y Región Oficina, las cuales tenían una correlación de 0,5 y al entrar al modelo daban valores en sentido contrario. Se realiza la regresión por separado y se deja la variable con mayor relación con respecto a la de respuesta, en este caso, Módulo. Esta variable está categorizada en Módulo G1 y Módulo G2 (ésta última queda libre). Módulo G1 está conformada por los módulos presentes en las regiones desde la VI a la XIV y la IV región. Esto quiere decir que en aquellas regiones, los costos son más bajos que en los módulos de las regiones del norte.

El modelo se ve coherente al analizar los valores de los coeficientes de cada variable independiente. Deuda Máxima corresponde a la máxima deuda de los últimos 6 meses de evaluación, por lo tanto, si es mayor el valor de esta deuda, esto repercute en que el valor del Costo Fijo sea mayor.

Atributo	Beta	Std. Error	T-stud	P-valor
Monto máquina	0.126	0.01	12.52	0.0
Si tiene vehículo	-0.168	0.079	-2.12	0.0387
Deuda máxima	0.038	0.0105	3.662	2.644E-4
Si tiene línea de crédito	0.072	0.011	6.810	1.020E-11
Ln línea de crédito disponible	0.0315	0.011	2.937	0.003593
Ln monto vehículo	0.331	0.0798	4.141	3.613E-5
Vivienda: Con deuda o arriendo	0.0734	0.0097	7.583	3.608E-14
Formalidad: Informal	-0.284	0.010	-26.64	0.0
Empleados: 2	0.333	0.010	32.537	0.0
Empleados: 3	0.403	0.010	39.20	0.0
Empleados: 4 o más	0.377	0.010	35.209	0.0
Cat rubro: G2	0.0655	0.010	6.508	7.936E-11
Modulo: G2	-0.11	0.00956	-11.49	0.0
Puntaje Ambiental	-0.027	0.013	-2.06	0.0448
Constante	11.55	0.009	1222.995	0.0

Indicador	MAE	MAPE	MSE	RMSE	RAE	RRSE	R^2
Valor modelo	1.15	5.78%	1.15	1.072	0.68	0.77	0.47

Tabla 23: Regresión lineal e Indicadores modelo Costo fijo sin Historia

Los indicadores del modelo dan cuenta que posee un error bastante más grande que el modelo con historia. No se toman muy en cuenta los valores de la suma total del error como MAE y MSE ya que además de que esté modelo no es tan preciso como el de clientes con historia, la cantidad de observaciones es mucho mayor y por lo tanto la suma de errores en valor será mayor. El MAPE es el doble que lo que es coherente con el R^2, el cual es la mitad que el modelo anterior y según este, con la regresión de la tabla 23, se estaría explicando solo el 47% de la varianza de los datos.

Luego de aplicarle exponencial a los resultados de la predicción, se calcula el indicador de sobrestimación y subestimación, el cual, también disminuye con respecto al modelo con historia. La cantidad de observaciones que se encuentran dentro del rango de 95% y 1,05% es muy baja. Al momento de otorgar créditos, es preferible que se sobrestimen los costos por sobre que se subestimen y en este caso la suma de la sobrestimación y la estimación en el rango es menor a la subestimación. Como se puede ver en la tabla 24, más de la mitad de los datos predicen un costo menor al real.

	Entrenamiento		Validación	
Estimación v/s Ventas	**N**	**%**	**N**	**%**
Sobre 5% estimación	3788	42,3%	1701	43%
Entre 95% y 105%	439	4,9%	185	4,67%
Bajo 5% estimación	4726	52,7%	2068	52,3%
Total	8953	100%	3954	100%

Tabla 24: Indicador MAPE para sobre y subestimación

4.3. MODELOS PARA MARGEN

Los últimos modelos en estimar son los de margen. Dada la distribución del margen, es que no es necesario aplicar ninguna transformación para normalizar. El margen proviene de la siguiente relación:

$$Margen = \frac{Ventas - Costo\ Variable}{Ventas} \quad \in [0,1]$$

4.3.1 Margen con Historia

Se estudian primero los estadísticos descriptivos más importantes del modelo para analizar su comportamiento. En la tabla 29 se puede ver que la moda es 1, lo que es extraño ya que implicaría que el costo variable de los segmento es 1 en muchos casos. Si se observa la ilustración 6, se puede ver que se ve claramente que existe la moda de 1 y además, se puede ver otra moda entre 0,4 y 0,5.

Mínimo	Moda	Mediana	Media	Max	Desv. Tip.
0,26	1	0,68	0,689	1	0,176

Tabla 25: Estadística descriptiva Margen con Historia

Los percentiles del margen, dan cuenta también de lo cargado que esta la distribución desde el 0,6 hacia el 1. Esto da cuenta de lo bajos que son los costos variables con respecto a las ventas en estos segmentos.

P5	P10	P25	P50	075	P90	P95
0,43	0,48	0,55	0,68	0,83	0,92	0,99

Tabla 26: Percentiles Margen con Historia

Luego se categorizan las variables cualitativas y se les aplica el test de independencia. No hay variables dependientes entre las categóricas. En el anexo F 5) se pueden ver las frecuencias y promedios del margen por cada una de las categorías en este tipo de variables. Por otro lado, al eliminar las variables con

baja correlación con el margen (punto 4.1) se va un 50% más de variables que en los otros modelos y en cuanto a variables continuas quedan solo 2 que son relevantes para Margen. Las estadísticas de éstas, se pueden ver en la tabla 31.

Variable	Mínimo	Mediana	Media	Máximo	Dev tip
Margen anterior	0	0,7	0,69	1	0,177
Costo variable bajo anterior	0	208.022,5	383.465,8	8.275.298	527.420

Tabla 27: Estadísticos descriptivos variables cuantitativas y continuas

Si bien Margen anterior y Costo variable anterior vienen de la relación:

$$Margen_{ant} = \frac{Ventas_{ant} - Costo\ Variable_{ant}}{Ventas_{ant}}$$

Dada la correlación entre ambas variables (-0,57) y luego de que entran al modelo con valores razonables, es decir, el Margen anterior entra de las primeras afectando positivamente al margen por estimar y viceversa, el Costo Variable Bajo Anterior entra restándole al margen, en menor magnitud, se opta por dejar ambas variables en el modelo. El Costo Variable Bajo Anterior corresponde al costo variable más bajo que posea los clientes asociados a un mismo rubro en el pivote anterior.

Atributo	Beta	Std error	T-student	P-valor
Margen anterior	0.14425	0.0011	124.68	0.0
Región oficina G2	0.00355	9.449E-4	3.760	1.791E-4
Región oficina G3	0.00370	9.479E-4	3.905	9.9E-5
Nuevo Antiguo vigente	0.00239	8.422E-4	2.842	0.0048
Ln costo variable bajo anterior	-0.0051	0.0010	-4.69	2.78E-6
Empleados: 1	0.0097	0.0010	8.962	0.0
Empleados: 2	0.00389	0.0010	3.7240	2.075E-4
Categoría rubro G2	0.00954	0.0013	7.256	4.358E-13
Categoría rubro G3	0.0035	0.0012	2.934	0.0036
Categoría rubro G4	0.0123	0.00121	10.204	0.0
CONSTANTE	0.689	8.4114E-4	820.160	0.0

Indicador	MAE	MAPE	MSE	RMSE	RAE	RRSE	R^2
Valor modelo	0,048	7,63%	0,006	0,075	0.326	0.439	0.807

Tabla 28: Regresión lineal e Indicadores Margen con Historia

El modelo, conformado en su mayoría por variables categóricas, tiene como variable dummy empleados, donde se ve que mientras más empleados menos se margina. Además incluye una variable que no aparece en los otros modelos. La

variable Nuevo Antiguo verifica el estado del tipo de cliente en la cartera vigente, tiene tres categorías: Antiguo, Nuevo antiguo y Nuevo vigente pero luego de la recategorización queda solo una variable dummy, la cual es 1 si el cliente es Nuevo vigente y 0 en caso contrario. Por otro lado, nuevamente entra la variable referente a la categoría de la microempresa. En este caso hay cuatro categorías donde la categoría libre corresponde a Alimentación y Amasandería, Fabricación de Productos Diversos y Mueblista, quienes marginan menos que los demás. Esto puede ser por poseer costos variables más altos lo que hace que el aporte al margen sea menor. Por otro lado la Categoría rubro 4, correspondiente a Belleza y Salud y Educación y Formación, es la que aporta un mayor valor al margen con mayor beta en comparación a los demás.

En cuanto a los indicadores del modelo, éste posee un coeficiente de determinación muy alto. Por otro lado, el MAE y el MSE dan valores muy bajos debido a que como el margen esta acotado entre 0 y 1, entonces nunca habrá una diferencia significativa entre la predicción y el valor real, por lo tanto no dicen mucho. El MAPE, da más grande que los otros modelos con historia. Este valor es más interpretable ya que si bien el rango de los valores del margen es menor que ventas y costos, la variación porcentual del pronostico con respecto al valor real es mayor. De todos modos un MAPE de un 7,63% está considerado bueno para el modelo.

Finalmente se calcula la sobre y subestimación. Más de la mitad de las observaciones, para la base de entrenamiento y validación, caen entre el 95% y 1,05% del valor real. Este valor es muy alto pero no quiere decir que el modelo sea tan bueno. Si se observa la ilustración 11, la cual compara el histograma de los valores reales versus los valores pronosticados, se ve que el modelo deja afuera a todas las personas que marginan 1, es decir, la moda.

	Entrenamiento		**Validación**	
Estimación v/s Ventas	**N**	**%**	**N**	**%**
Sobre 5% estimación	1457	25,9%	597	25%
Entre 95% y 105%	3047	54,2%	1269	53,1%
Bajo 5% estimación	1121	19,9%	523	21,9%
Total	5625	100%	2389	100%

Tabla 29: Indicador MAPE para sobre y subestimación

Se puede apreciar además la influencia de tener sólo variables categóricas en el modelo (dummies) lo que deja el modelo con estas “oscilaciones” en la variable de respuesta.

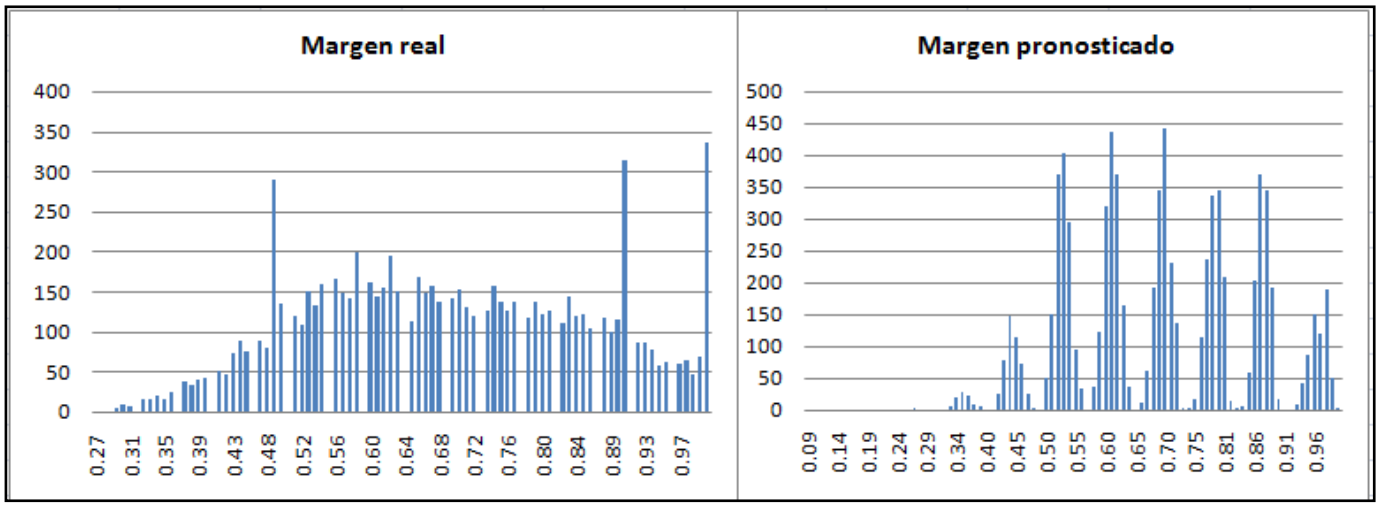


Ilustración 11: Comparación histogramas Margen real y pronosticado

4.3.2 Margen sin Historia

Se observan los estadísticos descriptivos principales del margen para clientes sin historia y se observa que tiene características muy similares a los clientes son historia. La moda es 1 nuevamente lo que quiere decir que probablemente el modelo no logre captar a todos estos clientes que no poseen costos variables.

Mínimo	Moda	Mediana	Media	Max	Dev. Tip.
0,14	1	0,69	0,695	1	0,177

Tabla 26: Estadística descriptiva Margen sin Historia

En los clientes sin historia hay porcentualmente más observaciones que marginan 1 que en clientes con historia. En Anexo E se puede ver el histograma de margen y se puede observar que hay cuatro modas más que sobresalen con respecto a los demás valores y están entre 0,5 y 0,9.

P5	P10	P25	P50	075	P90	P95
0,41	0,47	0,55	0,69	0,83	0,94	1

Tabla 27: Percentiles Margen sin Historia

Se categorizan las variables cualitativas y se le realiza el test de independencia. Hay dos variables que resultan ser dependientes, perfil riesgo y tipo cliente, pero finalmente estas no entran en el modelo. Por otro lado la variable Monto de Máquina, resulta tener una mejor correlación y coeficiente en el modelo de margen cuando está categorizada que cuando no. Esto se debe a que hay dos grupos estadísticamente distintos que se comportan mejor en el modelo como variable categórica ya que al dejarla como variable continúa se pierde esta diferencia entre grupos. En Anexo F se pueden ver las categorías que se incluyen en el modelo.

La tabla 28 presenta los resultados del modelo. Éste tiene la particularidad de presentar sólo variables categóricas en él. Se puede ver que mientras más Empleados hay, el margen es menor y ocurre lo mismo con la tenencia o no de vehículo (Permanencia vehículo). En cuanto a las categorías, G1 (Alimentación y Amasandería, Fabricación de productos y Mueblista) y G2 (Profesionales y Gestión Empresarial, Reciclaje y Recolector y Oficios) son los que hacen que el valor pronosticado para margen sea mayor, en especial para G2.

Atributo	Beta	Std error	T-student	P-valor
Si tiene vehículo vehículo	-0.0101	0.0015	-6.515	7.6E-11
Monto maquina menor a $1,3 mill	0.0129	0.0017	7.5	6.805E-14
2 empleados	-0.0181	0.0015	-11.91	0.0
3 o más empleados	-0.0389	0.0016	-23.83	0.0
Modulo G2	-0.01333	0.0014	-9.49	0.0
Cat rubro G3	0.0157	0.0019	8.229	2.22E-16
Cat rubro G2	0.0706	0.0021	33.98	0.0
Cat rubro G1	0.06304	0.0018	35.39	0.0
CONSTANTE	0.6957	0.0014	498.31	0.0

Indicador	MAE	MAPE	MSE	RMSE	RAE	RRSE	R^2
Valor modelo	0,025	21,1%	0,025	0,159	0.863	0.894	0.2

Tabla 28: Regresión lineal e Indicadores Margen sin Historia

Los indicadores del modelo afirman lo que se esperaba, que el modelo peores indicadores que Margen con Historia. El MAPE es mayor al 20% lo que es mucho más alto de lo que tendría un buen modelo y el coeficiente de determinación,R^2, es más bajo que en todos los modelos probados. Se cree que esto es por varias razones: Primero que nada porque no existe una variable que de noción de cual pudo haber sido el comportamiento del margen del cliente en el periodo anterior. Segundo, porque la distribución de la variable, bimodal con una modas de 0.48 y 1, no permite que el modelo prediga bien en ese rango. Finalmente, como última razón, se tiene que no hay variables independientes que predigan correctamente el margen en este caso.

Se calcula la sobre y subestimación de modelo. Si bien posee un MAPE más alto que Costo Fijo sin Historia, los resultados en este ítem son mejores. De todos modos, de acuerdo al modelo, se puede afirmar que un gran porcentaje de subestimaciones, corresponden al percentil más alto de margen, al cual no llega a estimar la regresión.

	Entrenamiento		Validación	
Estimación v/s Ventas	**N**	**%**	**N**	**%**
Sobre 5% estimación	3818	42,3%	1694	43,2%
Entre 0,95% y 105%	1330	14,7%	644	16,4%
Bajo 5% estimación	3874	43%	1579	40,3%
Total	9022	100%	3917	100%

Tabla 29: Indicador MAPE para sobre y subestimación

4.3.3 Modelo Alternativo

Como alternativa a este modelo, se analiza las distribuciones del Margen sin Historia entre segmentos con comportamiento similares. En las ilustraciones 12 y 13 se puede ver una notoria diferencia entre distintos segmentos, donde los primeros marginan mucho más que los segundos. Este tipo de fenómeno puede estar afectando los resultados y es por eso que se realiza un modelo que estima sólo a los segmentos M1, M2, M3, S4 y S5 correspondientes a Artesanado, Confección Calzado, Mueblista, Alimentación y Amasandería y Fabricación de Productos Diversos respectivamente.

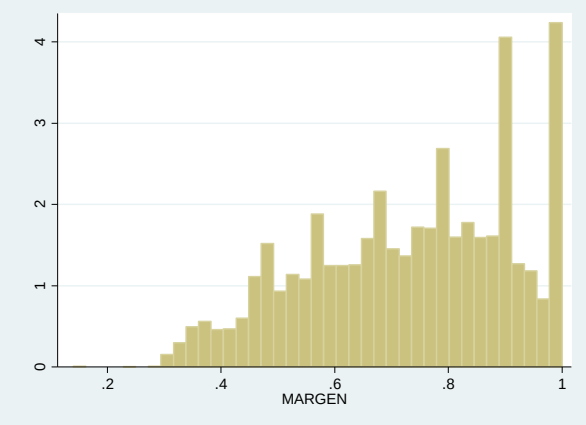


Ilustración 12 Histograma Margen Segmentos: M4, M5, S1, S2, S3

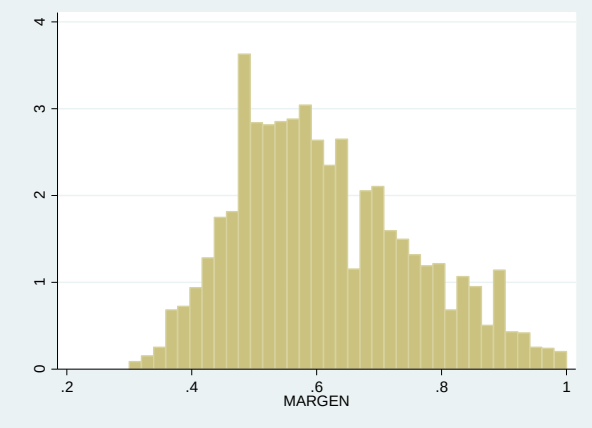


Ilustración 13: Histograma Margen Segmentos M1, M2, M3, S4, S5

Sin embargo los resultados que arroja son similares al modelo anterior, es una centésima mejor que el modelo original. El MAPE da un valor de un 20,9% y el MSE un 0,0248. El coeficiente de determinación tiene un valor de un 0,22. Si bien los resultados son mejores, no es un buen modelo y el hecho de tener que separar los segmentos, lo cual es más trabajo, hace que sea preferible quedarse con el modelo actual.

4.4. REDES NEURONALES

Los resultados obtenidos por los métodos lineales, si bien son rápidos y útiles para dar una aproximación general del pronóstico de ventas, costo fijo y margen, no son tan exactos y por lo tanto se investigan métodos no lineales para resolver este problema.

Se realiza una prueba de redes neuronales con el algoritmo “Perceptron Multicapa” y se aplica a Ventas sin Historia. Se realizan tres modelos moviendo los parámetros de Momentum, Aprendizaje y Training. Los resultados con respecto a la sobre y subestimación son los siguientes:

	Red Neuronal 1		**Red Neuronal 2**		**Red Neuronal 3**	
	N	%	N	%	N	%
Sobre 5% venta estimada	5904	45,7	5356	42,5	6109	47,3
entre venta estimada	1052	8,2	1120	8,7	1102	8,5
Bajo 5% venta estimada	5959	46,1	6439	49,8	5704	44,2

Tabla 30: Sobre y subestimación redes Neuronales

La regresión lineal califica un 8,1% de los datos entre 0,95% y 1,05%, lo cual no dista mucho de los resultados entregados por redes neuronales simples. Se pudo haber probado con un modelo más complejo pero el banco no quiere tener modelos muy complejos.

Se sugiere al banco separar las bases de datos por los distintos segmentos dadas las diferencias que presentan entre estos mismos. El caso más notorio es el del Margen. Para solucionar el problema del margen, se propone separar la base de datos tal como se hizo en el punto 4.3.3. mediante un árbol y luego que se calcule un modelo simple de redes neuronales para cada uno de estos.

5. CONCLUSIONES

El objetivo principal de este trabajo era construir seis modelos de regresión lineal que estimen ventas, costo fijo y margen para clientes con y sin historia de BancoEstado Microempresa de modo que se puedan calcular sus Estados Financieros y por lo tanto, su capacidad de pago al momento de otorgar créditos. Si bien hay muchos modelos que podrían aplicarse a este problema, se decide usar regresiones lineales por los bajos costos, simplicidad y rapidez que se tienen.

En primer lugar, se debió tratar todos los datos que BancoEstado otorgó para este trabajo. Aquí fue donde se eliminó el 10% de las observaciones para todas las bases de datos (ventas, costo fijo y margen) de clientes con y sin historia. Luego de realizar el filtro a las observaciones, se realizó uno a las variables. En un comienzo eran 136 variables por cada base y de acuerdo a la variable de respuesta, se fueron sacando aquellas variables que tenían una correlación muy baja con respecto a esta (correlación $\rho<|1|$), esta eliminación además se corroboraba utilizando el test T de mínimos cuadrados ordinarios.

Luego se debieron categorizar todas las variables cualitativas y algunas cuantitativas que lo requiriesen. Para ello se generaron N-1 variables dummy con respecto a las N categorías formadas. En el caso de margen con historia, se puede apreciar el particular efecto que tiene generar un modelo con casi todas sus variables dummy. Las oscilaciones que genera la presencia solo de variables dicotómicas hacen que no sea óptimo el modelo ya si bien mantiene la orientación de los valores pronosticados, se pierde un poco la robustez de los valores exactos.

Además de eliminar variables que eran muy insignificantes frente a la variable de respuesta, se tuvo que eliminar variables que estuvieran correlacionadas entre ellas para no generar problemas de multicolinealidad. En un comienzo ocurrió que se eliminaron todas aquellas variables que entre sí tuvieran una correlación mayor o igual a 0,7 dejando aquella que tuviera mayor ρ frente a la variable de respuesta correspondiente. El problema luego, fue que en un par de casos, como en ventas con historia, entraron dos variables que si bien no tenían una correlación muy alta (alrededor de 0,5) generaban problemas de colinealidad ya que deberían haber tenido un beta con el mismo signo, pero no ocurría así y se anulaban. Un ejemplo de esto fue para el caso de margen donde entraron al modelo las variables *Pertenencia de vehículo* y *Monto de vehículo,* la primera con signo negativo y la segunda con signo positivo. Para solucionar este problema, se dejó la que generara en un modelo lineal univariable, entre la variable de respuesta y la variable de prueba respectiva, el beta más alto.

Luego de generar los seis modelos y comparar los resultados entre si, se concluye que los modelos obtenidos para Ventas y Costo Fijo están dentro del rango de lo que esperaba el banco al compararlo con Comercio, el cual es un segmento en donde el banco ya ha comenzado a implementar TER Express. Se cree que si bien no son muy exactos los modelos de regresión lineal, el porcentaje

de pronósticos que están entre el 95% y 105% del valor real de la variable de respuesta, mas la subestimación (que es más preferible que la sobrestimación), son un porcentaje mayor al 60% lo cual es muy bueno y hace quedarse conforme con el trabajo al cumplir el objetivo central.

Por otro lado, se concluye que hay transformaciones que modelen en un 100% correctamente una variable que no se comporta de por sí como una normal. Los modelos con las transformaciones de Box Cox o logarítmicas tenían indicadores casi perfectos en Ventas y Costo Fijo. Sin embargo al aplicar la inversa de la transformación para obtener el valor real de Ventas y Costo Fijo, los indicadores cambiaban de valor, lo que no quieres decir que esté malo, sino, que al fin y al cabo el comportamiento de la variable de respuesta no se capta perfectamente.

En el caso de margen, se cree que está mal planteada la regresión lineal en una variable que se comporta entre 0 y 1. Se debe cambiar por un modelo probit o logit ya que por un lado, el procedimiento para calcular estos modelos es muy similar a la regresión lineal (a diferencia que trabaja con probabilidades) y por otro lado se asegura al banco que nunca se estimará un valor mayor a 1 para estos casos, ya que al trabajar con probabilidades, los valores siempre serán entre 0 y 1. Esta variable es la más compleja de modelar. Se recomienda al banco segmentarla por tipos de rubros y luego probar nuevamente modelos lineales y no lineales.

Finalmente, hay una gran conclusión con respecto a la importancia de conocer cualitativamente los segmentos y a los clientes. En un comienzo, los modelos se realizaron sin tener noción alguna de cómo era normal que se comportaran las microempresas en ciertas variables y por lo tanto los valores de prueba que dieron los modelos, no fueron los más óptimos. Es de vital importancia para un buen funcionamiento de estos modelos, ir actualizando los parámetros (betas) de acuerdo a como se vaya moviendo el mundo microempresarial constantemente. Hay estudios que indican cada cuanto tiempo es recomendable recalibrar los modelos.

- ## Recomendaciones

Analizar la posibilidad de incluir variables exógenas en el modelo. Estas pueden dar cuenta de cómo está la economía del país y por lo tanto podrían servir para realizar pronósticos. Por ejemplo, el PIB podría ser un buen indicador y podría verse reflejado en el crecimiento de las ventas de microempresarios. La tasa de desempleo sectorial, al influir directamente en ingresos y cantidad de empleados, también podría ser una opción de variable exógena para agregar, donde, se podría incluir un coeficiente por cada categoría de rubro.

Capacitar mejor a los ejecutivos del banco. Muchos de estos anotan los Costos Variables como Costos Fijos y eso hace que al momento de calcular el margen, este de 1. Esa es una buena medida para comenzar a aproximar la curva de Margen a una normal y eliminar la moda 1 que poseía en clientes con y sin historia.

Finalmente, se recomienda que si es que se va a implementar un modelo no lineal, que sea sencillo y no tan complejo. Es importante recordar que los resultados de estos modelos no son los que se llevar para calcular la capacidad de pago. Se elige el valor de resultado operacional mínimo entre los modelos y los cálculos hechos por ejecutivos con la información que les dan los clientes así que mientras no dependan de estas estimaciones en un 100%, no es tan relevante realizar modelos tan exactos. Distinto sería si el banco adoptara la postura de estimar el resultado operacional solo con modelos. En este caso, si convendría invertir en un modelo complejo.

6. BIBLIOGRAFÍA

[1] MORALES, L. YAÑEZ, A. Agosto 2007, "Microfinanzas en Chile, Resultados de la Encuesta de Colocaciones en Segmentos Microempresariales," Superintendencia de Bancos e Instituciones Financieras,Santiago. Pp.7-8

[2] GARAY, P. 2006, "CATASTRO Y DIAGNÓSTICO DE LAS MIPEs REGIÓN DE MAGALLANES Y ANTARTICA CHILENA," SERCOTEC Región de Magallanes y Antarica Chilena, Punta Arenas. Pp 31-33

[3] (2005) SOFOFA, Clasificación PYME. [Online]. HYPERLINK "http://www.sofofa.cl/sofofa/index.aspx?channel=4301" [Agosto 2013]

[4] LARRAIN, C. Mayo 2007, "BancoEstado Microcréditos: Lecciones de un modelo exitoso," CEPAL, Santiago de Chile. Pp. 11-13.

[5] BANCOESTADO MICROEMPRESA, 2012 "*Memoria anual"*, BancoEstado. Santiago. Pp. 32-34.

[6] Red para el Desarrollo de las Microfinanzas en Chile A.G., "Estado de las Mcrofinanzas en Chile 2012," Santiago, 2012.

[7] GERENCIA DE RIESGO BEME, 2012 "Desarrollo de Modelos TER EXPRESS," BancoEstado Microempresa (private communication) Santiago, Chile. Pp. 8-16.

[8] WEBB, A. 2002, *Statistical Pattern Recognition. 2ª ed*.: West Sussex, John Wiley & Sons, Ltd., p. 534.

[9] BIRON, M. Abril 2012 , "Desarrollo y Evaluación de Metodologías para la Aplicación de Regresiones Logísticas en Modelos de Comportamiento Bajo Supuesto de Independencia," Memoria de Ingenieria Civil, Universidad de Chile. Santiago, Chile,.

[10] ARAYA, C.. Archivos de Estadistica. [Online]. HYPERLINK "http://www.geocities.ws/estadistica/archivos/segmentacion.pdf" [Septiembre 2013]

[11] MORRIS H. DEGROOT AND MARK J. SCHERVISH, 2011 *Probability and Statistics*, 4th ed. Boston, Massachusetts: Addison Wesley.

[12] ZHANG, H., GUTIERREZ, H. 2010 *Teoria Estadistica: Aplicaciones y Métodos*, Primera edición ed. Bogotá, Colombia: Universidad Santo Tomás.

[13] GREENE, W. 2006 *Análisis Econométrico 3ra edición*. Nueva York: Prentice Hall.

[14] UNIVERSIDAD DE CORUÑA. Seleccion variables regresoras. [Online]. HYPERLINK "http://dm.udc.es/asignaturas/estadistica2/sec9_7.html" [Septiembre 2013]

[15] MORALES, P. Febrero 2008, "Correlación y Covarianza," *Estadística Aplicada a las Ciencias Sociales*, Universidad Pontificie Comillas, Madrid. Pp 15-16,.

[16] COVARRUBIAS, G., Abril 2012 "Construcción y Validación de una Metodología de Seguimiento para Modelos de Regresión Logística," Memoria de Ingenieria Civil .Santiago, Chile. Pp 4-30.

[17] ESTRATEGIA, DIARIO DE NEGOCIOS DE CHILE. [Online]. HYPERLINK "http://www.estrategia.cl/especiales/2012/ESP_LEASING_27092012.pdf" [Septiembre 2013]

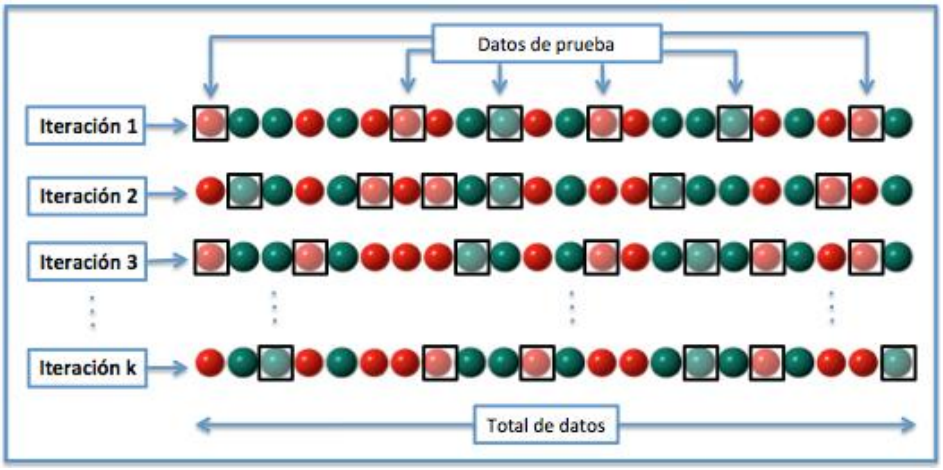
[18] HAN, J. KAMBER, M. 2005, Data Mining: Concepts and Techniques. 2ª ed. San Francisco: Morgan Kaufmann, Pp. 657-704.

7. ANEXOS

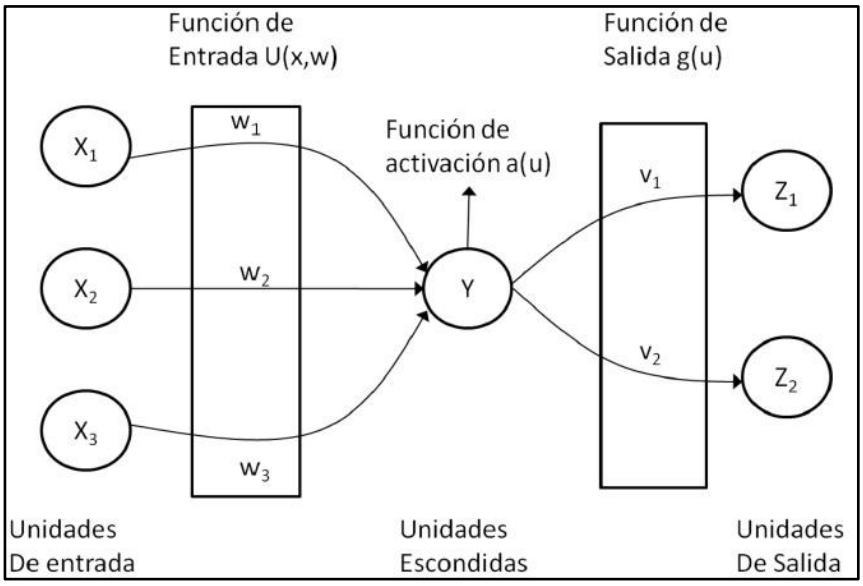
Anexo A: Caracterización de Instituciones Informantes

Institución	Tipo de Institución	Modalidad de créditos	Tipos de créditos
BancoEstado Microempresas	Banco público	Personales	Urbano – Agríc.
Santander Banefe	Banco privado	Personales	Urbano – Agríc.
Banco BCI Nova	Banco privado	Personales	Urbano
Indap	Institución pública	Personales	Agrícola
Banco Desarrollo de Scotiabank	Banco privado	Personales	Urbano
Oriencoop	Cooperativa	Personales	Urbano – Agríc.
Coopeuch	Cooperativa	Personales	Urbano
Emprende Microfinanzas	Sociedad Anónima	Personales	Urbano – Agríc.
Fondo Esperanza	ONG	Grupales	Urbano
CCAF de Los Andes	Caja de Compensación	Personales	Urbano – Agríc.
Corporación WWB – Finam	Corporación	Personales	Urbano
Fundación BanIgualdad	ONG	Grupales	Urbano
Fundación Contigo	ONG	Personales	Urbano
Fundación Kolping	ONG	Personales	Urbano
Fundación Crecer	ONG	Grupales	Urbano

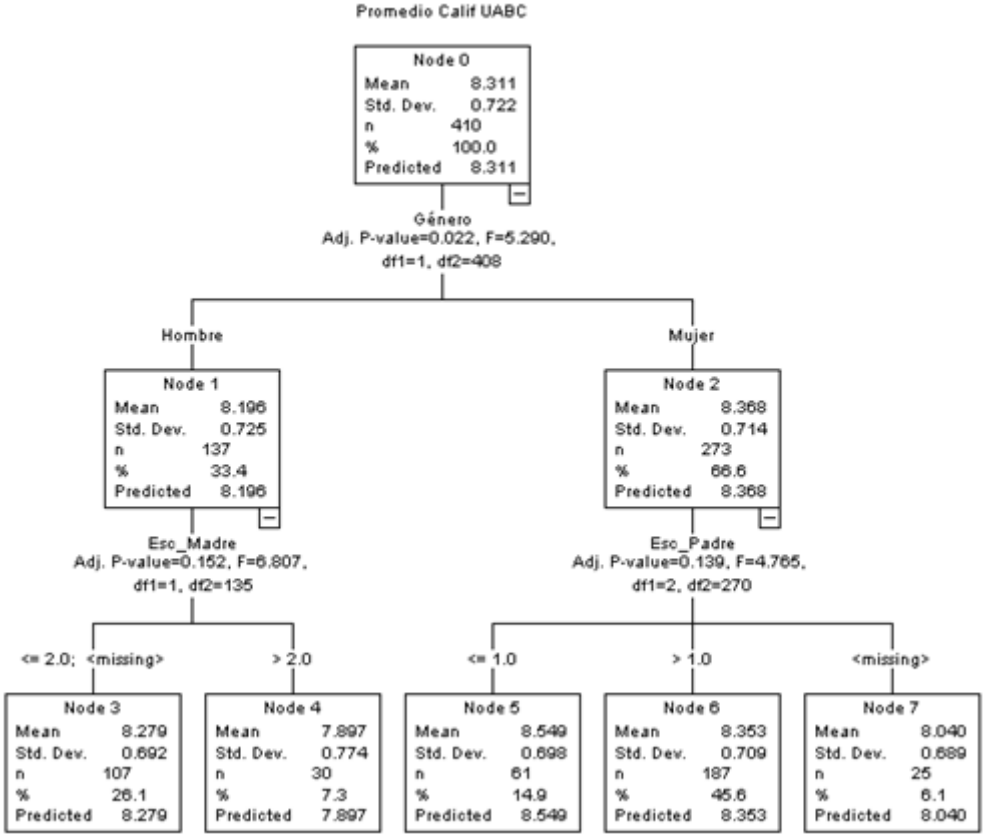
Anexo B: Validación cruzada aleatoria, método K-fold Cross



Anexo C: Redes Neuronales en Multicapa

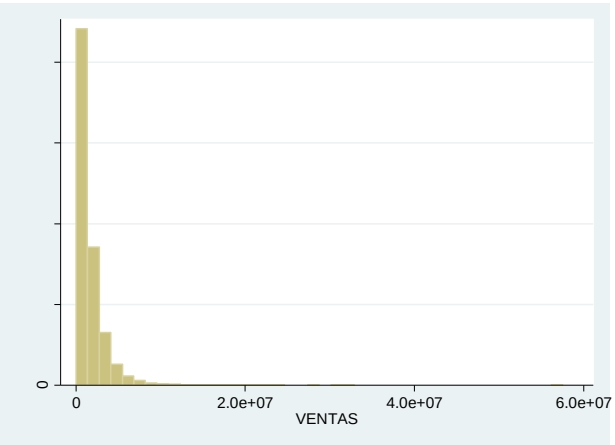


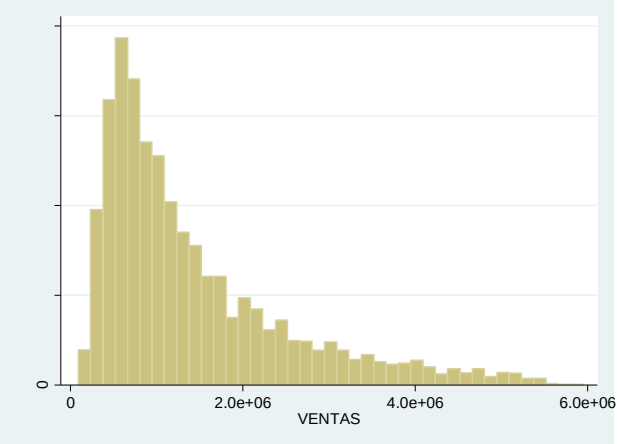
Anexo D: Algoritmo CHAID

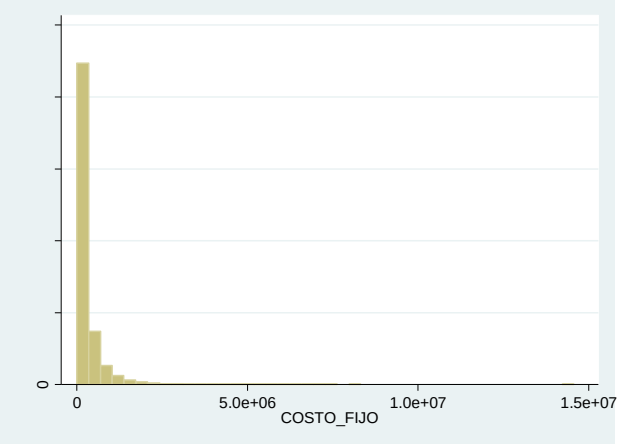


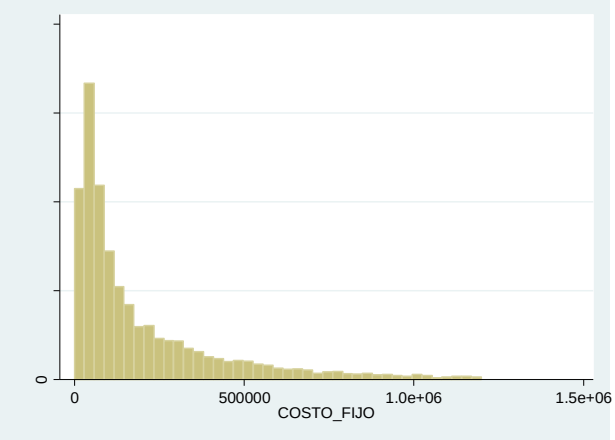
Anexo E: Histogramas de ventas, costo fijo y margen para clientes sin historia.

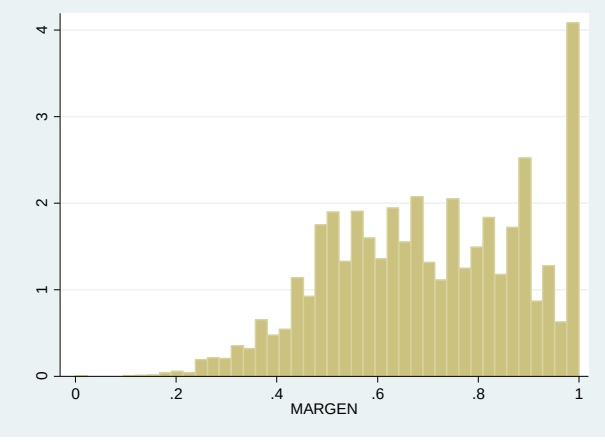
La primera columna es sin los filtros y la segunda columna luego de aplicar los filtros.

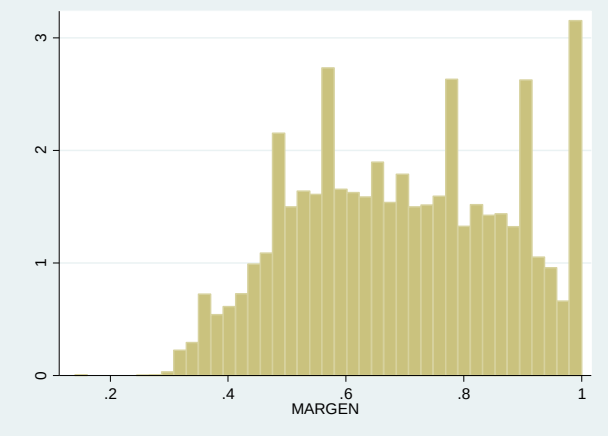












Anexo F Categorías Modelos

1) Categorías Ventas con Historia

Códigos Rubro	N	Porcentaje	Prom. Ventas
G1: M1, M4, M2	2022	26%	742.663,7
G2: S1, S2, S4, S3, S5, M3, M5 (libre)[3]	5769	74%	1.189.447,1

Vivienda	N	Porcentaje	Prom. Ventas
G1: Propia, Cedida o Allegado (libre)	6338	81,4%	990.533,5
G2: Propia con deuda o Arriendo	1453	18,6%	1.366.822,5

Formalidad	N	Porcentaje	Prom. Ventas
G1: Formal o Semiformal (libre)	5073	65,1%	1.349.168,8
G2: Informal	2718	34,9%	661.985,2

Empleados	N	Porcentaje	Prom. Ventas
G1: 1 empleado (libre)	4563	58,6%	762.226,2
G2: 2 empleados	1665	21,4%	1.269.329,3
G3: 3 o más empleados	1563	20,1%	2.206.679,7

Región Oficina	N	Porcentaje	Prom. Ventas
G1: X, IX, V, XIV, XII, XIII (libre)	3009	38,6%	961.258,8
G2: I, II, IV, VI, VII, VIII, XIII, XV	4778	61,4%	1.114.592,1

[3] La categoría que posee (libre) hace referencia a la cual no aparece en el modelo

Adultos (dummy)	N	Porcentaje	Prom. Ventas
G1: 1 adulto (libre)	5770	74,1%	1.137.108,4
G2: 2 o más adultos	2021	25,9%	842.390,4

Niños	N	Porcentaje	Prom. Ventas
G1: 0 niños (libre)	3966	50,9%	940.342
G2: 1 niño	1948	25%	1.102.398,8
G3: 2 o más niños	1877	24,1%	1.271.870,5

2) Ventas sin Historia:

Categoria Rubro	N	Porcentaje	Prom. Ventas
G1: S2, S3, S4, M5	7993	61,9%	1.260.475
G2: S1, S5, M3 (libre)	2643	20,5%	1.035.090,1
G3: M1, M2, M4	2279	17,6%	722.880,1

Vivienda	N	Porcentaje	Prom. Ventas
G1: Propia, Cedida o Allegado (libre)	10556	81,7%	1.041.319,5
G2: Propia con deuda o Arriendo	2359	18,3%	1.387.618,2

Formalidad	N	Porcentaje	Prom. Ventas
G1: Formal o Semiformal (libre)	8912	69%	642.420,5
G2: Informal	4003	31%	1.395.829.3

Empleados	N	Porcentaje	Prom. Ventas
G1: 1 empleado (libre)	7181	55,6%	755.396,9
G2: 2 empleados	2963	22,9%	1.269.329,3
G3: 3 o más empleados	2771	21,5%	2.206.679,7

Niños	N	Porcentaje	Prom. Ventas
G1: 0 niños	6428	49,8%	971.891

G2: 1 niño	3328	25,8%	1.360.005,5
G3: 2 o más niños	3159	24,5%	2.294.440.5

Adultos (dummy)	N	Porcentaje	Prom. Ventas
G1: 1 adulto (libre)	9788	75,8%	1.141.666,9
G2: 2 o más adultos	3127	24,2%	867.177,4

LCR (cuenta corr) dummy	N	Porcentaje	Prom. Ventas
No	10479	81,1%	1.244.425
Si	2436	18,9%	2.226.556

3) Costo Fijo con Historia

Empleados	N	Porcentaje	Prom. Costo fijo
G1: 1 empleado (libre)	4452	58,8%	59.218
G2: 2 empleados	1647	21,8%	150.842,7
G3: 3 o más empleados	1468	19,4%	300.738

Modulo	N	Porcentaje	Prom. Costo fijo
Mod g1: X R, V R Cord IX R Nort y Sur, XIV R, XI y XII R, IV R.	2608	34,5%	83.867
Mod g2: Otras (libre)	4959	65,5%	108.878,8

Códigos Rubro	N	Porcentaje	Prom. Costo fijo
G1: M1, M4, M2	2966	26%	742.663,7
G2: S1, S2, S4, S3, S5, M3, M5 (libre)	5601	74%	52.103,2

Vivienda	N	Porcentaje	Prom. Costo fijo
G1: Propia, Cedida o Allegado (libre)	6201	81,9%	91.125
G2: Propia con deuda o Arriendo	1366	18,1%	143.300,1

Formalidad	N	Porcentaje	Prom. Costo fijo
G1: Formal	2642	34,9%	149.491,3
G2: Informal y Semiformal (libre)	4925	65,1%	46.535,8

4) Costo Fijo sin Historia

Vivienda	N	Porcentaje	Prom. Costo fijo
G1: Propia, Cedida o Allegado (libre)	10561	81,8%	95.319
G2: Propia con deuda o Arriendo	2346	18,2%	152.511,2

Formalidad	N	Porcentaje	Prom. Costo fijo
G1: Formal	7627	40,9%	168.382,2
G2: Informal y Semiformal (libre)	5280	59,1%	51.584

Empleados	N	Porcentaje	Prom. Costo fijo
G1: 1 empleado (libre)	7198	55,8%	54.994
G2: 2 empleados	3006	23,3%	165.378
G3: 3 empleados	1479	11,5%	315.210
G4: 4 o más empleados	1224	9,5%	362.578,8

Región Oficina	N	Porcentaje	Prom. Costo fijo
G1: XIII, V, VII, III	5399	44,6%	112.869
G2: IX, X, XII, XIV	2443	20,2%	78.275
G3: VIII, IV,XI	3040	25,1%	92.317
G4: XV, II, I, VI	1212	10%	139.803

Códigos Rubro	N	Porcentaje	Prom. Costo fijo
G1: M5, S3, S2	7417	57,9%	130.743
G2: S1, S4, , S5, M1, M2, M3, M4 (libre)	5402	42,1%	73,129

5) Margen con Historia

Región Oficina	N	Porcentaje	Prom. Costo fijo
G1: VI, VII, VIII, X,XII, XV libre	2474	30,9%	0,667
G2: I, V,RM,IX,XIV	4458	55,7%	0,692
G3: II,III,IV,XII	1077	13,4%	0,735

Códigos Rubro	N	Porcentaje	Prom. Costo fijo
G1: S5, S4, M3 (libre)	1201	15%	0,57
G2: M5, M4, S3	3409	42,5%	0,721
G3: M1, M2	2020	25,2%	0,637
G4: S1, S2	1384	17,3%	0,793

Anexo G: Estadísticos Descriptivos Variables Cuantitativas sin historia

Ventas sin Historia:

Variable	Mínimo	Mediana	Media	Máximo	Dev tip
LC_disp	0	0	387,2	21.538	981,5
Puntaje SICA	196,7	906,2	875	994	102,3
ddvi	0	1.494	4.091,6	185.948	7716,3
Puntaje Ambiental	73,8	778,3	746.2	956.3	127.75
monto_mquina	0	1.034.000	2.373.860	9e+07	4.381.142

Costo fijo sin Historia:

Variable	Mínimo	Mediana	Media	Máximo	Dev tip
Dda_max	0	605	3.113,4	218.863	7466,1
LC_disp	0	0	382,2	21.047	1063,4
Monto vehiculo	0	0	2.040.621	9.9e+07	4.438.058
Puntaje Ambiental	62,5	817,3	778,1	968,76	138
monto_mquina	0	1.200.000	2.607.209	9.2e+07	5.016.935