



LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Fundación Universitaria los Libertadores.

**FORMULACIÓN DE UN MODELO PARA CALIFICAR LA PROBABILIDAD
DE FUGA DE CLIENTES VIGENTES EN UNA ENTIDAD DE MICROCRÉDITO.**

Presenta:

Javier Mauricio Castillo Galindo

Bogotá, Junio del 2019.



LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Fundación Universitaria los Libertadores.

**FORMULACIÓN DE UN MODELO PARA CALIFICAR LA PROBABILIDAD
DE FUGA DE CLIENTES VIGENTES EN UNA ENTIDAD DE MICROCRÉDITO.**

Trabajo para obtener el título de:

Especialista en Estadística Aplicada

Presenta:

Javier Mauricio Castillo Galindo

Asesor:

Mag. Heivar Yesid Rodríguez Pinzón

Bogotá, Junio del 2019.

Notas De Aceptación

Firma Presidente Del Jurado

Firma Del Jurado

Firma Del Jurado

Bogotá D.C, 22 de Junio de 2019.

Las directivas de la Fundación Universitaria Los
Libertadores, los jurados calificadores y el cuerpo docente
no son responsables por los criterios e ideas expuestas en el
presente documento. Estos corresponden únicamente
a los autores y a los resultados de su trabajo.

Contenido

1. Introducción.	11
2. Planteamiento del Problema.	13
2.1 Formulación del Problema.	18
2.2 Objetivos.	18
2.2.1 Objetivo General.	18
2.2.2 Objetivos Específicos.	18
2.3 Justificación.	19
3. Marco Conceptual.	21
3.1 Cliente Vigente.	21
3.2 Cliente Fugado o Cancelado.	21
3.3 Tasa de Fuga.	22
3.4 Microcrédito.	22
3.5 Análisis de correspondencias.	22
3.6 Modelo de Regresión logística Logit.	23
3.7 Métricas de Evaluación de la Clasificación.	24
3.7.1 Matriz de Confusión.	24
3.7.2 Sensibilidad.	25
3.7.3 Especificidad.	25
3.7.4 Tasa Acierto.	26
3.7.5 Mala Clasificación.	26
3.7.6 Curva de ROC.	26

4. Marco Metodológico.	28
4.1 Programas.	28
4.2 Población.	28
4.3 Variables.	29
4.4 Metodología Aplicada.	31
4.4.1 Depuración de la Base.	31
4.4.2 Análisis de Correspondencias.	31
4.4.3 Aplicación de modelo de clasificación.	31
4.4.4 Evaluación del modelo.	31
4.4.5 Interpretación de Resultados.	32
5. Desarrollo y Resultados.	33
5.1 Depuración de la Base.	33
5.1.1 Selección de variables y/o clientes a excluir.	33
5.1.2 Validación de datos nulos.	33
5.1.3 Transformación de variables, numéricas a categóricas, categóricas a factores numéricos.	34
5.2 Análisis de Correspondencias.	34
5.2.1 Construcción de graficas de correspondencias multivariado.	34
5.3 Aplicación de modelo de clasificación.	38
5.3.1 Selección de muestra para construcción y prueba del modelo.	38
5.3.2 Modelo de regresión Logit.	39
5.4. Evaluación del modelo.	40
5.4.1 Matriz de confusión.	40

5.4.2 Curva ROC.....	42
6. Conclusiones.....	43
7. Recomendaciones.....	44
8. Bibliografía.....	45

Lista de Tablas

Tabla 1 Ejemplo Tabla de confusión	24
Tabla 2 Distribución de datos de análisis.....	28
Tabla 3 Variables cuantitativas disponibles para análisis.....	29
Tabla 4 Variables cualitativas disponibles para análisis.....	30
Tabla 5 Conteo de Clientes Entrenamiento.....	38
Tabla 6 Conteo de Clientes Validación.....	38
Tabla 7 tabla se confusión de validación	40

Lista de Figuras

Figura 1 Distribución porcentual de los microestablecimientos por rangos de personal.....	13
Figura 2 Distribución porcentual del personal ocupado por tipo de contratación.	14
Figura 3 Cifras entidades Microfinancieras Marzo 2019.....	15
Figura 4 Grafico de ROC.....	27
Figura 5 Porcentaje Pagado (Porc_Pago).....	35
Figura 6 Número de Días en Mora (Num_Dias_Mora).....	35
Figura 7 Nota	36
Figura 8 Tasa Crédito Vigente (Max_Tasa)	36
Figura 9 Antigüedad Crédito Vigente (Antigüedad_Cred_Vig).....	37
Figura 10 Edad del Cliente (Edad).....	37
Figura 11 Salida del Modelo Logit En R Studio.....	39
Figura 12 Grafica ROC Validación	42

FORMULACIÓN DE UN MODELO QUE CALIFIQUE LA PROBABILIDAD DE FUGA DE CLIENTES VIGENTES EN UNA ENTIDAD DE MICROREDITO.

Javier Mauricio Castillo Galindo¹

Resumen.

Prevenir la fuga de clientes es una de las prioridades de las entidades financieras y aún más importante para las entidades de microfinanciamiento, debido al costo que representa la vinculación de un cliente; Por esto tener una herramienta que permita calificar la probabilidad de fuga de los clientes vigentes del activo es de mucha ayuda pues así se puede decidir en qué clientes enfocar los esfuerzos de retención.

Este trabajo se inicia con un análisis descriptivo de las variables cualitativas a partir de un análisis de correspondencias encontrando, por ejemplo, que a pesar de lo que se pensaba la tasa no está fuertemente relacionada con la fuga de clientes.

Paso seguido se ajusta un modelo de regresión Logit (Modelo de respuesta binaria) el que permite calificar la probabilidad de fuga de los clientes a partir de unas variables de entrada que según su relación con la fuga permiten ajustar esta probabilidad, para este caso las variables que se identificaron significativas para la deserción fueron:

Número de créditos Vigentes, Porcentaje Pago, saldo en mora mayor, días de mora, nota, antigüedad del crédito vigente, plazo del crédito, numero de microseguros, tasa, reestructurado, numero de productos vigentes, Estado del Ahorro, Saldo del ahorro, Número de pólizas vigentes, Edad, Ámbito, Estado civil, Estrato, SOW y Segmento de relación.

Palabras Claves: Microfinanciamiento, Retención, Fuga, Análisis de correspondencia, Regresión Logit, SOW (Share of wallet).

¹ Ingeniero Industrial, Universidad Católica de Colombia, jmcastillog@ulibertadores.edu.co

1. Introducción

El presente trabajo busca aportar una herramienta estadística que permita calificar la probabilidad de fuga de clientes de una entidad microfinanciera a partir de un conjunto de variables disponibles las cuales serán la base del estudio.

La fuga de clientes es una de las principales preocupaciones que tienen las entidades financieras pues el prepago de créditos o la no renovación de estos hace que el número de clientes vigentes disminuya afectando su crecimiento y a su vez la rentabilidad del negocio, pues no es solo la falta de ingresos por intereses o por costo de transacciones lo que disminuye sino también el costo que conlleva la recuperación de clientes la que hace que la operación se más costosa y por consiguiente menos rentable.

Pero ¿cómo prevenir la fuga de clientes?, ¿cómo enfatizar los esfuerzos comerciales en la retención sino se tiene un foco específico?, se podría dar la instrucción de atender a todos los clientes después de cierto porcentaje pagado de sus obligaciones, pero ¿Los clientes que prepagan? Y si se omite esto ¿a cuántos clientes se debería atender?, la operación comercial se agotaría y por consiguiente se encarecería, es aquí donde la estadística aporta valor pues permite calificar la probabilidad de fuga de los clientes basándose en la historia y comportamientos previos para poder enfocar sus esfuerzos de retención.

Para empezar con la comprensión del comportamiento de los clientes fugados se realiza un análisis de correspondencias que permite describir la relación de las variables cualitativas contra la fuga o no de clientes; esto para tener indicios de que variables se pueden incluir en el modelo a realizar.

Para este caso y debido al tipo de respuesta que se espera es Binaria (Fuga, No Fuga) se formulara un modelo de regresión logística Logit, el cual permite identificar las variables más representativas en la fuga de los clientes calificándolas según su aporte al modelo y estableciendo una probabilidad a cada cliente según sus variables y categorías propias.

Paso seguido se realizaran las validaciones al modelo comprobando su capacidad de calificar la fuga de clientes con una base de validación que se creó específicamente para este fin y cuyos datos no están incluidos dentro de la base utilizada para la creación del modelo.

Con todo lo anterior se busca establecer un modelo inicial que permita realizar pruebas piloto en las oficinas ajustando las políticas de retención a estos clientes en específico.

2. Planteamiento del Problema

En Colombia no es solo fundamental pensar en las grandes empresas que generan empleos directos apalancados en grandes capitales económicos, sino es necesario enfocarse en los microempresarios que a partir de pequeños capitales, negocios tradicionales y muchas ganas, buscan establecer negocios que les permita tener un ingreso para suplir sus necesidades.

Para este caso cabe resaltar que cuando se habla de microempresarios, se tiene en cuenta las personas que tienen negocios de bajo capital, por ejemplo: ventas ambulantes (Frutas, Jugos, dulces, etc.), ventas de abarrotes, ventas por catálogo, emboladores, pequeños agricultores, entre otros; Estos normalmente son negocios de una sola persona y que generan bajos ingresos.

Figura 1 Distribución porcentual de los microestablecimientos por rangos de personal.

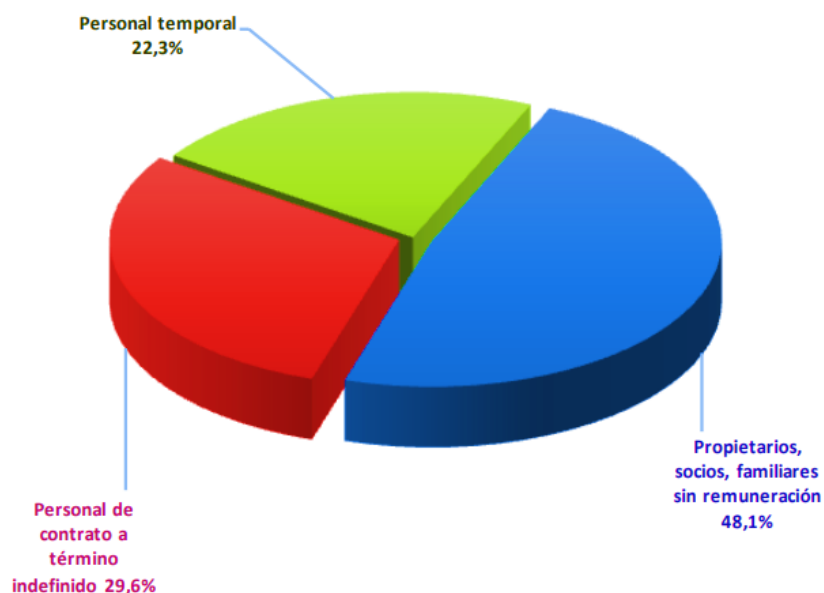


(DANE, 2016)

Como lo muestra el Dane en su informe de la Encuesta Panel De Microestablecimientos realizada entre el 2015 y el 2016, en la que resalta que el 71.3% de los microestablecimientos

normalmente están compuestos por máximo dos trabajadores y a su vez muestra que en el 48.1% de estos establecimientos los empleados no reciben salario, pues son propietarios, socios y/o familiares los que laboran ahí y solo reciben las ganancias que deje el día.

Figura 2 Distribución porcentual del personal ocupado por tipo de contratación.

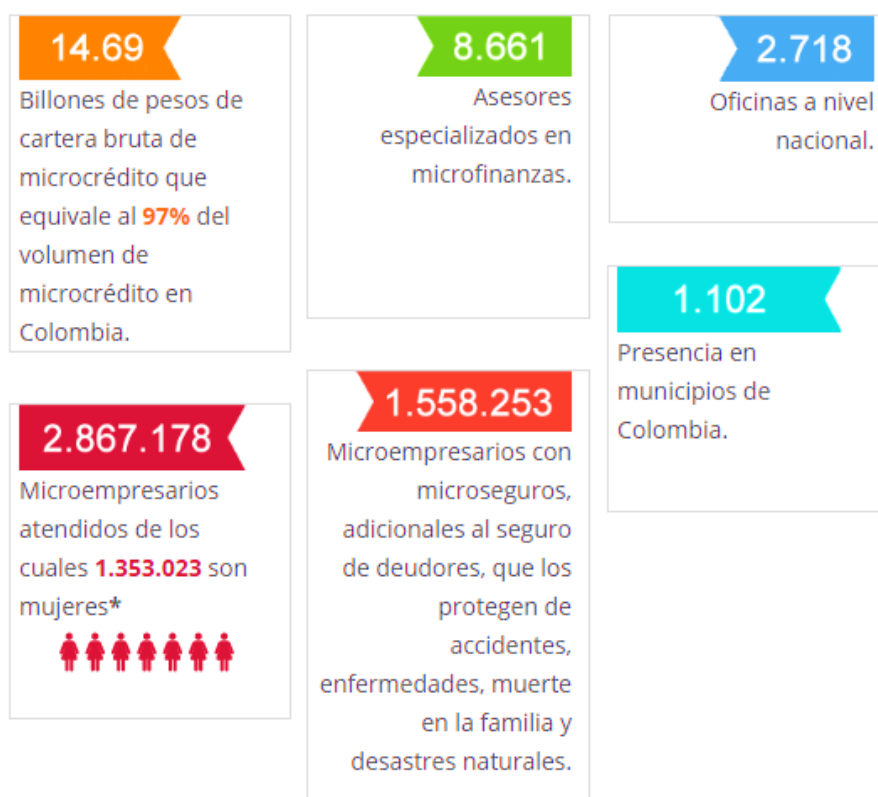


(DANE, 2016)

Al ser personas que tiene bajos capitales y que en su mayoría no tienen experiencia crediticia, el acceso a capital de inversión es muy limitado tanto que muchas veces optan por utilizar el medio más fácil pero menos conveniente que ofrecen grupos ilegales en Colombia y es la financiación a través del “gota a gota”, la revista semana explica este tipo de préstamos como: “Te entregan dinero sin pedirte garantías, a cambio de que se lo devuelvas en cuotas diarias y con intereses altos, por lo general tienes 20 o 30 días para devolver el monto más un interés del 10% al 40%.” (Semana, 2019). La anterior siendo una alternativa demasiado costosa y peligrosa, es en muchos casos la única opción.

Dado lo anterior en Colombia varias fundaciones y bancos ofrecen créditos que buscan apoyar a los microempresarios ofreciéndoles productos financieros para su negocio y dándoles la oportunidad de ingresar al sector financiero sin las mismas condiciones que normalmente exigen las grandes entidades bancarias, tanto así que para marzo del 2019 Asomicrofinanzas reporta que las entidades microfinancieras atienden a más 2.867.178 microempresarios colombianos, con una presencia en 1.102 municipios a nivel nacional, véase figura3.

Figura 3 Cifras entidades Microfinancieras Marzo 2019.



(Asomicrofinanzas, 2019)

Aun cuando la finalidad es apoyar al pequeño empresario, estos créditos tienen tasas muy altas comparadas a las de la banca tradicional, esto se debe a varios factores:

- Riesgo: El perfil del cliente y su inestabilidad económica hace que el riesgo sea mayor y por consecuencia requiere asumir mayores costos por castigos y provisiones.

- Costos de vinculación: el objetivo fundamental del crédito es apoyar a personas que tengan negocios propios, y para poder establecer un monto adecuado no solo a las necesidades del negocio sino, a su vez, a sus capacidades, el ejecutivo debe realizar una o varias visitas en las que realice el levantamiento de la información del negocio (balance, inventario), referencias comerciales y vecinales, entre otros; Por lo cual el costo de vinculación, estudio del crédito y desembolso es alto, esto sin tener en cuenta las zonas rurales donde los tiempos de recorrido son mayores y la productividad del ejecutivo menor.

- Educación financiera: al ser fundaciones en su mayoría o bancos que vienen de una fundación, su objetivo principal es apoyar al microempresario brindándole herramientas que le permitan mejorar su negocio apoyándolos con talleres continuos de educación financiera lo cual también genera un costo a la operación del negocio.

- Costo de Fondeo²: Por su tamaño y su falta de productos disponibles de ahorro, la captación de dinero es un problema para estas entidades por lo que deben acudir a otras entidades financieras que les realizan préstamos a tasas crediticias más costosas, lo que se ve reflejado en las tasas de crédito hacia los clientes y en la utilidad de la entidad.

Por lo anterior para estas entidades el costo perdido que representa la fuga de un cliente es muy alto, pues el costo de recuperación de un cliente fugado es casi igual al de vinculación de un cliente nuevo, y es muy inferior comparándolo al de un cliente que renueva su obligación

² Costo (Intereses) que implica el conseguir fondos para poder prestar dinero a sus clientes.

pues los requisitos debido a su continuidad son menores y hacen que los desembolsos sean más ágiles y menos costosos.

No solo esto hace que la labor de estas entidades sea complejo, pues al vincular clientes que inicialmente otras organizaciones rechazan por su historial crediticio, se les da la oportunidad de crecer y lamentablemente cuando esto ocurre, los clientes “Buenos”³ son atraídos por otras entidades que tienen diversidad de servicios, tasas más competitivas vinculándolos y haciéndolos que terminen su relación con la entidad de microcrédito.

Además de esto, al ser fundaciones o entidades bancarias pequeñas las que ofrecen este tipo de créditos y según el perfil ya antes mencionado del cliente, el principal objetivo es reducir costos y hacer pocas inversiones en temas diferentes a la operación básica del negocio debido a esto, el análisis estadístico que se ha realizado en este tipo de entidades es muy poco, por el costo que este representa, por lo que al remitirse a las referencias bibliográficas no se encuentra mucha información específica de este perfil de cliente o modelos aplicados específicamente a este tipo de negocio, se encuentra comúnmente, modelos ajustados a entidades financieras comerciales (Oppenheim, 2017) (Miranda, 2014), tesis sobre el impacto de las microfinanzas en desarrollo económico de los países (Sanz, 2016), propuestas para mejorar las microfinanzas (Torres, 2016), entre otras.

Para este trabajo se busca aportar a la entidad financiera, que por términos de privacidad se referencia en el trabajo como “Banco A”, una herramienta para calificar la posibilidad de cancelación de crédito de sus clientes vigentes, para generar estrategias de atención diferenciales que disminuyan la fuga de sus clientes de crédito.

³ Clientes “Buenos”; clientes que a pesar de sus dificultades son cumplidos con sus obligaciones con la entidad.

2.1 Formulación del Problema

Con el fin de mejorar las estrategias de retención y dirigirlas a un mejor grupo objetivo de clientes vigentes disminuyendo los esfuerzos operativos de la red comercial de la entidad A, es necesario calificar o estimar su posibilidad de fuga de ahí que sea necesario determinar

¿Cuáles clientes tienen mayor probabilidad de fuga?

2.2 Objetivos

2.2.1 Objetivo General

- Formular un modelo estadístico que permita calificar la probabilidad de fuga de los clientes vigentes del crédito de la entidad financiera Banco A.

2.2.2 Objetivos Específicos

- Identificar las variables más relevantes en la fuga de clientes, a partir de un análisis de correspondencias de las variables cualitativas disponibles de los clientes vigentes del crédito.

- Formular un modelo estadístico que califique la probabilidad de fuga de los clientes vigentes de crédito de la entidad financiera Banco A.

2.3 Justificación

La fuga para las entidades financieras es uno de los principales problemas que enfrentan en el día a día, y más para este tipo de entidades en las que el costo de vinculación de un cliente es tan alto, debido a todos los procesos que conllevan su vinculación; por esto es fundamental maximizar los esfuerzos que se realizan al mantener vinculados a sus clientes pero a su vez minimizando el costo operativo de esta labor.

Pero ¿cómo prevenir la fuga de clientes?, ¿cómo enfatizar los esfuerzos comerciales en la retención sino se tiene un foco específico?, se podría dar la instrucción de atender a todos los clientes después de cierto porcentaje pagado de sus obligaciones, pero ¿Los clientes que prepagan? Y si se omite esto ¿a cuántos clientes se debería atender?, la operación comercial se agotaría y por consiguiente se encarecería, es aquí donde la estadística aporta valor pues permite calificar la probabilidad de fuga de los clientes basándose en la historia y comportamientos previos para poder enfocar sus esfuerzos de retención; además permite:

- Identificar las variables que afectan la fuga o no de un cliente vigente.
- Calificar la posibilidad de fuga de un cliente vigente.

A pesar de que se encuentra información disponible de modelos estadísticos de este tipo, no se encuentran trabajos aplicados a este nicho en particular de clientes, pues la mayoría de entidades microfinancieras invierten poco en este tipo de estudios por su falta de recursos e información.

Por lo anterior, este trabajo busca aportar conocimientos a una entidad financiera de microcrédito sobre sus clientes vigentes del activo, aportando valor a la predicción de fuga de estos en determinado tiempo.

3. Marco Conceptual

Para contextualizar lo desarrollado en este trabajo, se presentaran las definiciones más importantes utilizadas, se detallaran las variables disponibles de la base de datos y sus categorías y la metodología implementada para el planteamiento del modelo de calificación de fuga, explicando el paso a paso ejecutado.

3.1 Cliente Vigente.

Se refiere a aquella persona natural que actualmente tiene una obligación financiera (Microcrédito) con saldo ante la entidad.

3.2 Cliente Fugado o Cancelado.

Se refiere a aquella persona natural que paga en su totalidad el saldo e intereses de sus obligaciones financieras (Microcrédito) con la entidad.

Para este caso de análisis el cliente que cancela sus obligaciones pero aún posee un producto de ahorro (Cuenta de Ahorro - CDT), se incluye dentro del concepto de cliente fugado.

Es importante resaltar que según la definición anterior, los clientes que por sus días de mora fueron castigados por la entidad no se incluyen dentro del análisis ni en el conteo de clientes fugados.

3.3 Tasa de Fuga.

Este es un indicador muy utilizado por las empresas para calcular la proporción de clientes que cancelan sus servicios en un periodo de tiempo determinado comparándolo contra la cantidad de clientes iniciales del periodo, normalmente el periodo de cálculo es mensual.

$$\text{Tasa de Fuga} = \frac{\text{Clientes cancelados}}{\text{Clientes Inicio del Periodo}} * 100$$

El principal objetivo es garantizar un equilibrio entre los clientes fugados y los nuevos ingresados, y de esta manera poder establecer metas que garanticen un crecimiento del negocio.

3.4 Microcrédito.

Es un crédito de baja cuantía, y que se ofrece exclusivamente a personas con un negocio constituido pues busca su principal objetivo es apoyar a los microempresarios, como lo expresa las naciones unidas: “Es una pequeña cantidad de dinero prestada a un cliente por un banco u otra institución. El microcrédito puede ofrecerse, a menudo sin garantía, a una persona o mediante préstamos colectivos” (Unidas, 2004)

3.5 Análisis de correspondencias.

El análisis de correspondencias es una técnica descriptiva que por medio de análisis grafico nos permite comparar y agrupar variables categóricas identificando patrones y relaciones, “El análisis de correspondencia (CA) es una extensión del análisis de componentes principales adecuado para explorar relaciones entre variables cualitativas (o datos categóricos). Al igual que

el análisis de componentes principales, proporciona una solución para resumir y visualizar el conjunto de datos en gráficos de dos dimensiones.” (KASSAMBARA, 2017)

De los resultados que brinda el análisis de correspondencias se utilizara el grafico simétrico, el cual permite establecer relación entre las variables cualitativas comparadas, esta relación se asume a partir de la distancia o cercanía de las categorías de cada variable.

Se utiliza esta técnica debido a que se desea identificar relaciones entre las variables disponibles y la fuga o no del cliente y así dar indicios para la implementación del modelo.

3.6 Modelo de Regresión logística Logit

Debido a que el problema que se plantea en este trabajo es calificar y/o identificar que clientes tienen mayor probabilidad de fuga (Respuesta Binaria, 0 No fuga, 1 Fuga), se implementara un modelo de regresión logística ajustándolo a las variables disponibles.

Se utilizara el modelo Logit dado que permite obtener estimaciones probabilísticas de un suceso (Fuga o no), además de ayudar a identificar las variables que determinan estas probabilidades y establecer un peso de ellas contra la variable objetivo, permitiendo describir el comportamiento del suceso ante la variación de cada variable, la forma reducida del modelo de regresión Logit es:

$$\text{Prob}[Y_i = 1] = \frac{1}{1 + e^{-X_i'\beta}}$$

El modelo Logit, busca clasificar a los individuos en alguno de los dos grupos posibles (0 - 1) a partir de las relaciones que crea entre las variables explicativas y la variable objetivo,

además su resultado da un vector con parámetros cuyos coeficientes califican las variables explicativas según su importancia para el modelo.

3.7 Métricas de Evaluación de la Clasificación

Después de determinar el modelo final es necesario evaluar el desempeño del mismo, pues a pesar que se determinen variables significativas no representa que el modelo califique de manera correcta, para esto se tienen varias herramientas, las cuales se aplicaran en la valoración del modelo resultante, y así decidir si el resultado que este calcula es el esperado.

3.7.1 Matriz de Confusión

La matriz de confusión es la herramienta que se aplica para validar el comportamiento de los modelos de calificación binaria, y determinar si el modelo resultante se ajusta a un mínimo deseable.

Tras la generación del modelo este se aplica a la base de validación, para evaluar su comportamiento y paso seguido, la matriz cuenta las calificaciones reales vs las estimadas generando una matriz del tipo:

Tabla 1 Ejemplo Tabla de confusión

		Real	
		0	1
Modelo	0	TN	FN
	1	FP	TP

Fuente: Elaboración Propia.

Dónde:

TN - TP: Valores correctamente clasificados (Modelo se ajusta a la realidad).

FP: El modelo califica la fuga a un cliente que no se fugó (Error Tipo 1 – Falsa Alarma)

FN: El modelo deja de calificar a un cliente fugado (Error Tipo 2)

Para la validación de la tabla de confusión se definieron las siguientes métricas (SWETS, 1996)

3.7.2 Sensibilidad

La sensibilidad o recall mide la capacidad del modelo de clasificar correctamente los casos positivos:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3.7.3 Especificidad

La especificidad o **specificity** mide la capacidad del modelo de clasificar correctamente los casos negativos:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

3.7.4 Tasa de acierto

La tasa de acierto o accuracy, indica el ajuste del modelo con la calificación total del modelo, (Proporción de casos bien clasificados):

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Poblacion}}$$

3.7.5 Mala Clasificación

La mala calificación, indica el ajuste del modelo con la calificación total del modelo, (Proporción de casos mal clasificados):

$$\text{Mala Clasificación} = \frac{\text{FP} + \text{FN}}{\text{Total Poblacion}}$$

3.7.6 Curva de ROC

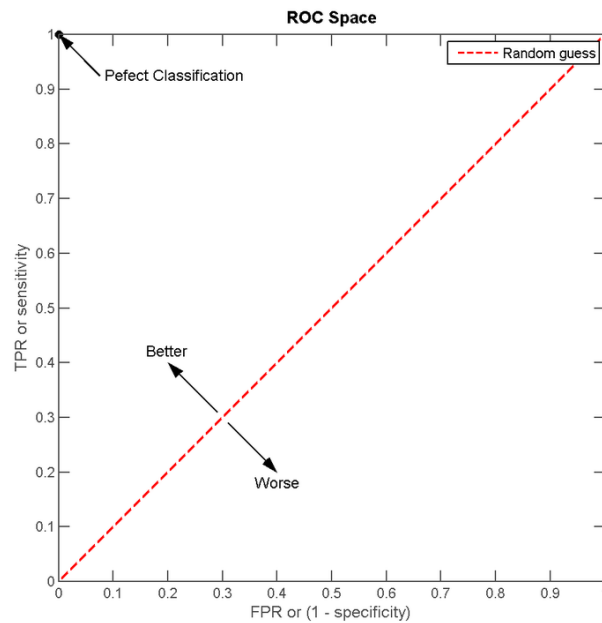
La Curva de ROC un gráfico representa el rendimiento de un modelo de clasificación, para este caso binario, Esta grafica vincula la tasa de verdaderos positivos (Sensibilidad) versus la tasa de falsos positivos (google, 2018):

Tasa de falsos positivos:

$$\text{Tasa de Falsos Positivos} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

El análisis del grafico es sencillo pues entre mayor sea el área captada bajo la curva mayor será su validez.

Figura 4 Grafico de ROC



(google, 2018)

El AUC es otro indicador que permite comprender el buen funcionamiento del modelo y expresa el área que se abarca bajo la curva ROC, su valor varía entre 0 y 1 y se puede interpretar según los siguientes (google, 2018) :

De 0.5 a 0.6 Teste es malo

De 0.6 a 0.75 Test es Regular

De 0.75 a 0.9 Test es Bueno

De 0.9 a 0.97 Test es Muy Bueno

De 0.97 a 1 Test es Excelente

4. Marco Metodológico

4.1 Programas.

Los programas utilizados para el desarrollo de la tesis fueron:

Programa estadístico R versión 3.6.0, entorno estadístico de R Studio, con sus paquetes "ROCR", "Rcmdr", "FactoMineR", "RcmdrPlugin.FactoMineR".

4.2 Población

El análisis se realizara sobre la base de clientes vigentes del crédito a un corte en específico del año 2019, de cada cliente se tienen unas variables propias que lo describen, además de la marca donde se especifica si el cliente se fugó al final del periodo (Variable Objetivo); La proporción de los clientes vigentes vs desertados para el análisis es la siguiente:

Tabla 2 Distribución de datos de análisis.

Situación	Categoría	Clientes	% Clientes
Clientes Vigentes	0	270.719	97,4%
Clientes Cancelados	1	7.151	2,6%
Total De Clientes		277.870	

Fuente: Elaboración Propia.

4.3 Variables

Para el análisis de los clientes vigentes de la entidad A, se tiene acceso a variables cualitativas y cuantitativas de cada cliente, en la siguiente tabla se presenta a detalle cada variable y sus categorías:

Tabla 3 Variables cuantitativas disponibles para análisis.

Variable	Descripción
NRO_CREDS_VIG	Numero créditos vigentes del cliente.
MONTO_CREDS_VIG	Monto inicial desembolsado de los créditos vigentes del cliente.
PORC_PAGO	Porcentaje pagado de sus créditos.
SALDO_MMAYOR	Saldo del cliente en caso de tener mas de 30 días de mora.
SALDO_MMENOR	Saldo del cliente en caso de tener entre 1 a 30 días de mora.
NUM_DIAS_MORA	Numero de días en mora del cliente con la entidad.
MIN_NOTA_VIG	Calificación interna del cliente que se asocia a su comportamiento de pago (1 Malo - 5 Excelente).
ANTIGUEDAD_CRED_VIG	Meses de antigüedad del crédito vigente.
MAX_PLAZO	Días de plazo del crédito vigente.
NUM_MICRO_SEGURO	Numero de microcréditos vigentes financiados.
MAX_TASA	Tasa máxima de los créditos vigentes del cliente.
NRO_PRODUCTOS	Nro. de productos vigentes (Ahorro - cdt - seguro).
SALDO_AHORRO	Saldo de la cuenta de ahorros vigente antes de cancelar su crédito.
NRO_POLIZAS_VIG	Numero de pólizas de seguros vigentes del cliente.
EDAD	Edad del cliente.
ESTRATO	Estrato Socioeconómico asociado al lugar de negocio o residencia del cliente.
INGRESO_VTAS	Dinero que le ingresa mensualmente al cliente por las ventas realizadas en su negocio.
EXCEDENTE_MENSUAL	Excedente de dinero del cliente después de restar a sus ingresos todas sus obligaciones y gastos.
MAX_MORA_12M	Máximo numero de días en mora del cliente con la entidad en los ultimo 12 meses.
TOTAL_CREDS_HISTO	Numero de créditos históricos que ha tenido el cliente con la entidad.

Fuente: Elaboración propia

Tabla 4 Variables cualitativas disponibles para análisis.

Variable	Descripción	Categorías	Categorías en R
YY	Variable Objetivo, cliente cancela o no su crédito.	No_Cancelo	0
		Cancelo	1
TIPO_CLIENTE	Marca si el cliente tiene o no un producto de ahorro aparte del crédito.	Excl_Activo	1
		Mixto	2
GENERO	Sexo del Cliente.	Femenino	1
		Masculino	2
AMBITO	Ámbito donde el cliente desarrolla su negocio.	Rural	1
		Urbano	2
REESTRUCTURADO_VIG	Marca si el cliente tiene o no un crédito reestructurado.	Reestructurado	1
		No_Reestructurado	0
EST_AHORRO	Estado de sus cuentas de ahorros.	No_Tiene	0
		Activo	1
		Activo_Otro	2
		Inactivo	3
ESTADO_CIVIL	Estado civil del cliente.	Casado	1
		Divorciado	2
		Soltero	3
		Unión_Libre	4
		Viudo	5
TIPO_VIVIENDA	Clasificación según la propiedad del inmueble donde vive el cliente.	Alquilada	1
		Asignada_Por_La_Empresa	2
		Familiar	3
		Propia	4
SECTOR	Categoría que se asigna según el tipo de actividad económica que realiza en su negocio.	Agropecuario	1
		Comercio	2
		Servicios	3
		Transformación	4
R_SOW_TOTAL	Es una clasificación que se le da al cliente, dependiendo el saldo de sus obligaciones con la entidad	Compartido	1
		Exclusivo	2
SEG_RELACION	Segmento asignado al cliente según clasificación propia de la entidad.	A.Rel	1
		B.Rel	2
		C.Rel	3
		D.Rel	4
SEG_CICLO	Segmento asignado al cliente según clasificación propia de la entidad.	A.Cic	1
		B.Cic	2
		C.Cic	3
		D.Cic	4

Fuente: Elaboración propia

Las anteriores servirán para el desarrollo de los objetivos y fueron depuradas previamente para incluir información de valor que no genere inconsistencias al realizar los análisis.

4.4 Metodología Aplicada

4.4.1 Depuración de la Base

- Selección de variables y/o clientes a excluir
- Validación de datos nulos
- Transformación de variables, numéricas a categóricas, categóricas a factores numéricos

4.4.2 Análisis de Correspondencias

- Construcción de graficas de correspondencias multivariado

4.4.3 Aplicación de modelo de clasificación

- Selección de muestra para construcción y prueba del modelo (Training - Validation)
- Modelo de regresión Logit

4.4.4 Evaluación del modelo

- Matriz de confusión
- Curva ROC
- Validación de especificidad sensibilidad

4.4.5 Interpretación de Resultados

- Variables que influyen en la fuga de clientes en la entidad A

- Interpretación del modelo final

5. Desarrollo y Resultados

5.1 Depuración de la Base

5.1.1 Selección de variables y/o clientes a excluir.

Antes de iniciar el análisis y de utilizar la data disponible es necesario realizar depuración de clientes que no son necesarios para el análisis y por el contrario pueden alterar el resultado del mismo; en esto caso se filtró:

- Al conocer que los créditos desembolsados tienen un plazo mínimo de seis meses, se excluyen los clientes con un desembolso menor a dos meses pues al validar los clientes cancelados con esta condición son casi nulos.

- Se excluyen los clientes con marca de cancelado debido a castigo pues este tipo de clientes no son el objetivo del análisis.

- Se excluyen los clientes con mora mayor a 90 días de atraso debido a que en su mayoría terminan en castigo, pues por el perfil de cliente de la entidad A, las labores de reestructuración son fuertes entre los 30 a 60 días de mora, para atrasos mayores la posibilidad de recuperación es muy baja.

5.1.2 Validación de datos nulos

Se revisa cada variable para garantizar que no se tengan datos nulos o faltantes, en los casos con estas condiciones se revisó la data de la entidad y se completó según correspondiera.

5.1.3 Transformación de variables, numéricas a categóricas, categóricas a factores numéricos.

Para el análisis de correspondencias no solo se incluyeron las variables categóricas disponibles, sino además, se ranguearon las variables numéricas para incluirlas en el análisis.

Para la construcción del modelo, por recomendación del tutor, las variables cualitativas fueron transformadas a categorías representadas por números e ingresadas a R, como factores, pues el manejo de estas variables para la aplicación es más ágil de esta manera.

5.2 Análisis de Correspondencias

5.2.1 Construcción de graficas de correspondencias multivariado.

Con la ayuda del complemento de R, R-Comander se realizó el análisis de correspondencias multivariado, de cada una de las variables disponibles contra la variable Objetivo (YY – 1 Cancelo o 0 No), para determinar que variables pueden aportar a la identificación de clientes fugados.

Las variables más representativas según este análisis fueron:

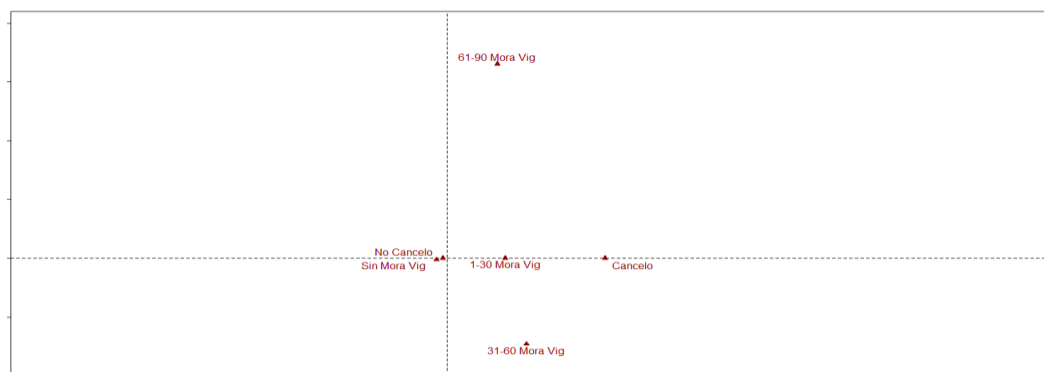
Figura 5 Porcentaje Pagado (Porc_Pago).



Fuente: Elaboración Propia.

Se observa que los clientes cancelados están más relacionados con aquellos clientes que han pagado más del 90% de sus obligaciones vigentes.

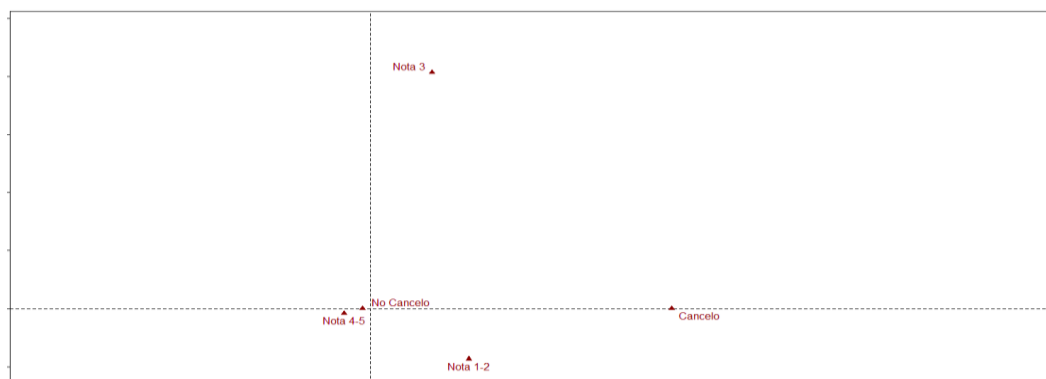
Figura 6 Número de Días en Mora (Num_Dias_Mora).



Fuente: Elaboración Propia.

Los clientes No Cancelados no están relacionas fuertemente contra los clientes que tienen algún tipo de mora, podría dar indicios de relación contra la fuga.

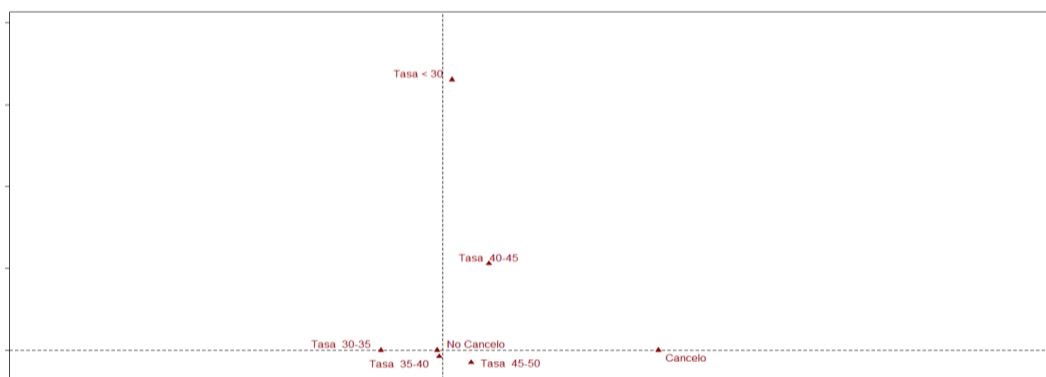
Figura 7 Nota



Fuente: Elaboración Propia.

Se esperaría que los clientes de mejor nota fuesen los de mayor relación contra la fuga, pues son más apetecidos por la competencia y por esto las ofertas percibidas son mayores, pero el grafico nos indica lo contrario.

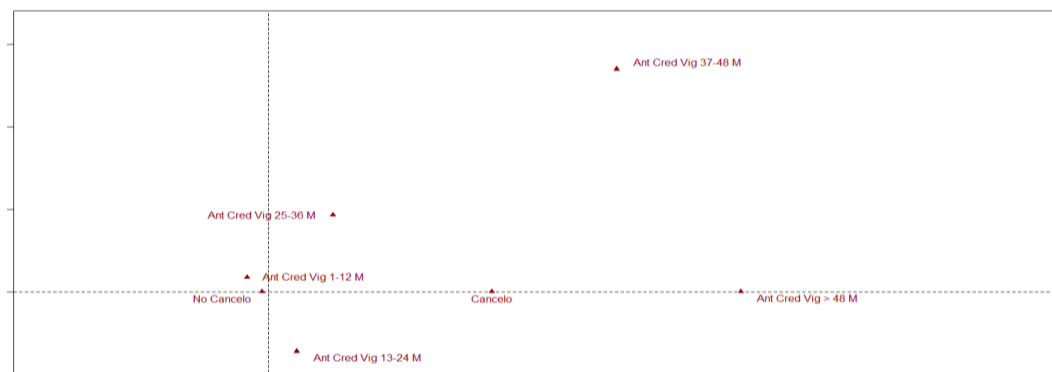
Figura 8 Tasa Crédito Vigente (Max_Tasa)



Fuente: Elaboración Propia.

Se esperaba que esta variable fuese más determinante al momento de evaluar la fuga, pues se pensaba que la tasa era una de las principales razones de los clientes para cancelar su crédito con la entidad.

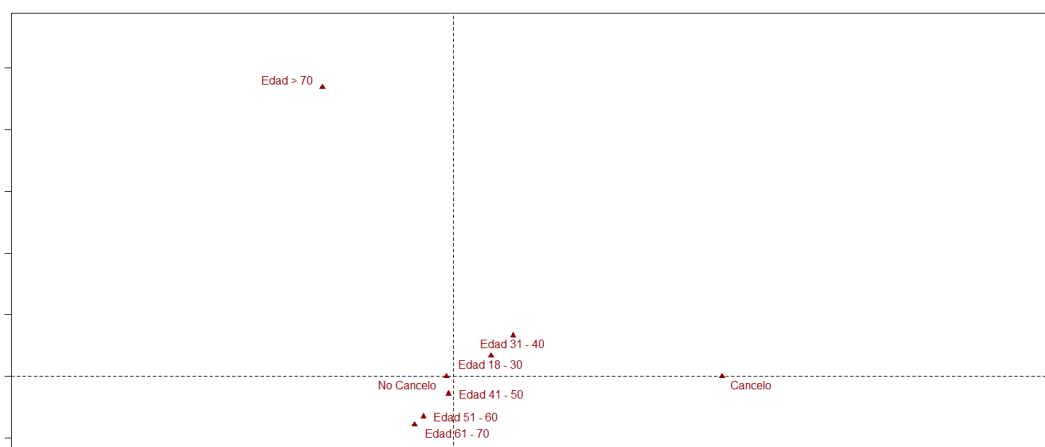
Figura 9 Antigüedad Crédito Vigente (Antigüedad_Cred_Vig)



Fuente: Elaboración Propia.

Los clientes más antiguos tienen mayor relación con la fuga, esto puede estar relacionado con el agotamiento del cliente ante la obligación de cancelar su cuota mensual.

Figura 10 Edad del Cliente (Edad)



Fuente: Elaboración Propia.

Se esperaba que las personas jóvenes estuviesen mayores relacionadas con la fuga, pero este análisis refuta ese preconcepto.

Aparte de las anteriores graficas se realizaron análisis con más variables las cuales no fueron representativas para la comparación de clientes fugados.

5.3 Aplicación de modelo de clasificación

5.3.1 Selección de muestra para construcción y prueba del modelo

Antes de iniciar es necesario generar dos bases, una base de entrenamiento (Training) y una base de validación (Validation), estas bases serán generadas a partir de la instrucción de R “Sample”.

La base de Entrenamiento (Training) es con la que se construirá el modelo Logit y corresponde al 75% de los datos de la base, tras la instrucción “Sample” la base generada es la siguiente:

Tabla 5 Conteo de Clientes Entrenamiento.

Situación	Categoría	Clientes	% Clientes
Clientes Vigentes	0	203.058	97,4%
Clientes Cancelados	1	5.344	2,6%
Total Clientes		208.402	

Fuente: Elaboración Propia.

Con el restante 25% de los datos se construye la base de validación (Validation)

Tabla 6 Conteo de Clientes Validación.

Situación	Categoría	Clientes	% Clientes
Clientes Vigentes	0	67.661	32,5%
Clientes Cancelados	1	1.807	0,9%
Total Clientes		69.468	

Fuente: Elaboración Propia.

Con esta se realizara la validación del funcionamiento del modelo.

5.3.2 Modelo de regresión Logit.

Con la base de entrenamiento y las variables disponibles se construye el modelo de regresión Logit, hasta ajustarlo y solo incluir las variables que sean representativas, con lo que se obtuvo el siguiente resultado:

Figura 11 Salida del Modelo Logit En R Studio

```
Call:
glm(formula = as.factor(YY) ~ NRO_CREDS_VIG + PORC_PAGO + SALDO_MMAYOR +
  NUM_DIAS_MORA + MIN_NOTA_VIG + ANTIGUEDAD_CRED_VIG + MAX_PLAZO +
  NUM_MICRO_SEGURO + MAX_TASA + REESTRUCTURADO_VIG + NRO_PRODUCTOS +
  EST_AHORRO + SALDO_AHORRO + NRO_POLIZAS_VIG + EDAD + AMBITO +
  ESTADO_CIVIL + ESTRATO + R_SOW_BMIA_TOTAL + SEG_RELACION,
  family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9277  -0.1707  -0.0859  -0.0564   4.8233

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.800e+00  2.222e-01 -30.606 < 2e-16 ***
NRO_CREDS_VIG -9.363e-01  1.054e-01  -8.880 < 2e-16 ***
PORC_PAGO      7.682e+00  1.167e-01  65.848 < 2e-16 ***
SALDO_MMAYOR  -7.048e-07  1.322e-07  -5.330  9.80e-08 ***
NUM_DIAS_MORA  1.338e-02  1.376e-03   9.722 < 2e-16 ***
MIN_NOTA_VIG2  8.029e-02  6.979e-02   1.151  0.249930
MIN_NOTA_VIG3  1.629e-01  7.486e-02   2.176  0.029587 *
MIN_NOTA_VIG4  1.091e-01  7.183e-02   1.519  0.128826
MIN_NOTA_VIG5  1.799e-01  5.858e-02   3.072  0.002128 **
ANTIGUEDAD_CRED_VIG -8.007e-02  4.679e-03 -17.112 < 2e-16 ***
MAX_PLAZO      7.856e-04  1.251e-04   6.281  3.37e-10 ***
NUM_MICRO_SEGURO  4.663e-01  4.683e-02   9.956 < 2e-16 ***
MAX_TASA       9.429e-01  2.522e-01   3.738  0.000185 ***
REESTRUCTURADO_VIG1 -3.235e-01  1.082e-01  -2.991  0.002783 **
NRO_PRODUCTOS  -3.852e-01  4.189e-02  -9.195 < 2e-16 ***
EST_AHORRO1    5.899e-01  8.534e-02   6.912  4.76e-12 ***
EST_AHORRO2    5.022e-01  1.756e-01   2.860  0.004230 **
EST_AHORRO3    2.259e-01  7.598e-02   2.974  0.002941 **
SALDO_AHORRO   3.786e-08  9.242e-09   4.097  4.19e-05 ***
NRO_POLIZAS_VIG -5.304e-01  3.881e-02 -13.669 < 2e-16 ***
EDAD           -6.292e-03  1.247e-03  -5.045  4.54e-07 ***
AMBITO2        7.040e-02  3.352e-02   2.100  0.035722 *
ESTADO_CIVIL2 -4.760e-02  5.588e-02  -0.852  0.394264
ESTADO_CIVIL3 -1.380e-01  4.448e-02  -3.103  0.001915 **
ESTADO_CIVIL4 -8.371e-02  3.870e-02  -2.163  0.030518 *
ESTADO_CIVIL5 -4.134e-02  8.577e-02  -0.482  0.629770
ESTRATO2       7.314e-02  3.504e-02   2.087  0.036850 *
ESTRATO3       1.462e-01  5.143e-02   2.843  0.004472 **
ESTRATO4       1.998e-01  1.532e-01   1.304  0.192140
ESTRATO5       6.707e-02  3.662e-01   0.183  0.854664
ESTRATO6      -1.274e+00  1.050e+00  -1.214  0.224755
R_SOW_BMIA_TOTAL2 -2.108e-01  3.137e-02  -6.721  1.80e-11 ***
SEG_RELACION2  8.622e-01  5.535e-02  15.576 < 2e-16 ***
SEG_RELACION3  7.871e-01  6.208e-02  12.679 < 2e-16 ***
SEG_RELACION4  4.147e-01  5.680e-02   7.300  2.87e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 49705  on 208401  degrees of freedom
Residual deviance: 34641  on 208367  degrees of freedom
AIC: 34711
```

Fuente: Elaboración Propia.

Se determinaron que las variables significativas en la fuga de clientes de la entidad A, son:

Número de créditos Vigentes, Porcentaje Pago, saldo en mora mayor, días de mora, nota, antigüedad del crédito vigente, plazo del crédito, numero de microseguros, tasa, reestructurado, numero de productos vigentes, Estado del Ahorro, Saldo del ahorro, Número de pólizas vigentes, Edad, Ámbito, Estado civil, Estrato, SOW y Segmento de relación.

Con el anterior modelo se realizaron las pruebas para validar su ajuste, con la data que se generó para este fin (Validation).

5.4. Evaluación del modelo

5.4.1 Matriz de confusión

Se genera la tabla de confusión con el siguiente resultado:

Tabla 7 tabla se confusión de validación

		Real	
		0	1
Modelo	0	63.918	533
	1	3.778	1.239

Fuente: Elaboración Propia.

Y se realizan las pruebas especificadas:

- Sensibilidad (Recall)

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{1239}{1239 + 533} = \mathbf{0,70}$$

- Especificidad (Specificity)

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{63918}{63918 + 3778} = \mathbf{0.944}$$

Con lo anterior se puede concluir que el modelo es bueno clasificando los clientes fugados pues acierta en un 70% de los casos, pero es mucho más efectivo clasificando los clientes que no cancelan sus obligaciones pues acierta en un 90% de los casos.

- Tasa de acierto (Accuracy)

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Poblacion}} = \frac{63918 + 1239}{69468} = \mathbf{0.938}$$

- Mala Clasificación

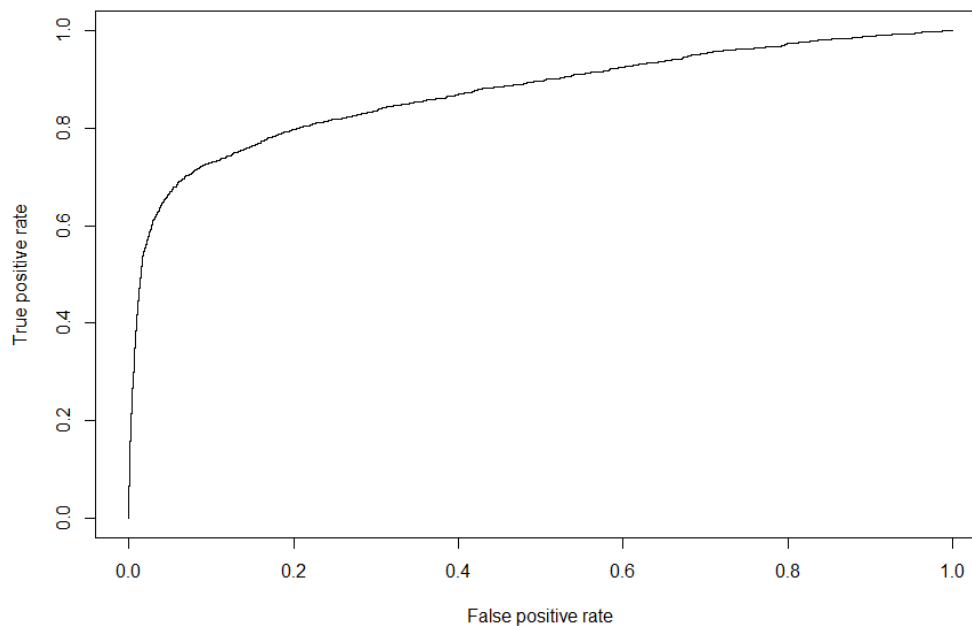
$$\text{Mala Clasificación} = \frac{\text{FP} + \text{FN}}{\text{Total Poblacion}} = \frac{3778 + 533}{69468} = \mathbf{0,062}$$

En general los indicadores permiten dar un parte de ajuste al modelo.

5.4.2 Curva ROC

Con los resultados obtenidos se ejecuta la gráfica ROC, la cual genera el siguiente gráfico:

Figura 12 Grafica ROC Validación



Fuente: Elaboración Propia.

Además se calcula el área bajo la curva:

AUC: 0.8722196

Con estos dos resultados, y según las definiciones iniciales se concluye que el modelo es bueno calificando la probabilidad de fuga de clientes de la entidad.

6. Conclusiones

- A pesar del peso que tienen los clientes no fugados en la base 97.4% contra 2.6% de los clientes fugados, fue fundamental la realización del filtrado inicial de la base en la que se evitaron incluir clientes que por sus características no eran viables de cancelar sus obligaciones, esto permitió que a pesar de las diferencias en las proporciones se pudiese ajustar un modelo.

- En el análisis de correspondencias se pudieron identificar variables fuertemente relacionadas con la fuga, a pesar de que parece obvio los clientes que han pagado el 90% de sus obligaciones son los más relacionados a la fuga, pero esto no es suficiente pues si solo la entidad se guía por este indicador podría estar omitiendo clientes que prepagan sus obligaciones, o podrían caer en la atención a una base de clientes muy grande y que no necesariamente van a desertar dejando de atender a clientes que por sus características adicionales si son potenciales de fuga.

- Las variables identificadas como significativas para el modelo eran las esperadas y reflejan en algunos casos lo encontrado en el análisis de correspondencias.

- En general los indicadores de validación aplicados a la tabla de confusión permiten dar tranquilidad al ajuste del modelo, la sensibilidad (Predicción de Fugados) es el indicador de menor calificación pero se ajusta a lo esperado para esta primera exploración.

7. Recomendaciones

- Las variables con las que se ajustó el modelo están disponibles dentro de las bases consolidadas por la entidad, se recomienda acudir a entidades externas que posean información adicional del cliente y de su comportamiento y así poderlas incluir al modelo para intentar realizar un ajuste adicional que mejore la sensibilidad; Dentro de las variables que se esperan incluir:

- Comportamiento del cliente con otras entidades (Calificación).
- Créditos vigentes con el sector financiero.
- Endeudamiento Total
- Capacidad de Pago

Con estas variables se esperaría poder replantear el modelo.

8. Bibliografía

- Asomicrofinanzas. (1 de Abril de 2019). asomicrofinanzas. Obtenido de asomicrofinanzas:
<https://asomicrofinanzas.com.co/cifras/>
- DANE. (2016). Encuesta Panel De Microestablecimientos Octubre 2015 - Septiembre 2016. Bogotá: DANE.
- google. (28 de 9 de 2018). Aprendizaje Automatico. Obtenido de
<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419>
- KASSAMBARA, A. (2017). Practical Guide to Principal Component Methods in R. Alboukadel Kassambara.
- Miranda, M. A. (2014). "Identificación De Un Modelo Explicativo De Retención De Clientes Con Riesgo De Fuga Para Una Entidad Bancaria Aplicando Regresión Logística Y Árboles De Clasificación . Lima, Peru.
- Oppenheim, D. A. (31 de 7 de 2017). Modelos de Churn Bancarios con R. Modelos de Churn Bancarios con R. Chile.
- Sanz, J. A. (2016). El impacto de las microfinanzas sobre el crecimiento económico de los países en vías de desarrollo. El impacto de las microfinanzas sobre el crecimiento económico de los países en vías de desarrollo. Valladolid, España.
- Semana. (2019). Qué son los préstamos "gota a gota" que grupos criminales de Colombia exportan al resto de América Latina. Semana, 5-7.
- SWETS, J. A. (1996). Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. Mahwah, NJ: Lawrence Erlbaum Associates.
- Torres, I. J. (2016). Estrategias para el mejoramiento de los programas de microfinanzas en Colombia. Bogotá, Colombia.
- Unidas, D. d. (2004). ¿Como pueden 100 dolares modificar una economia? New York: Naciones Unidas .