MEMORIA METODOLÓGICA

Modelo de Ingreso Clientes
Banco Falabella Perú

Agosto 2019

Índice

DESCRIPCIÓN DEL DOCUMENTO

Título	**Modelo de Ingreso Clientes**
Descripción del documento	Ficha metodológica del desarrollo del modelo de ingresos para clientes del Banco Falabella, en base a información interna del Banco y del sistema financiero.
Responsable de la documentación	**Banco Falabella Perú**
Fecha de aprobación	

ACTUALIZACIONES Y MODIFICACIONES

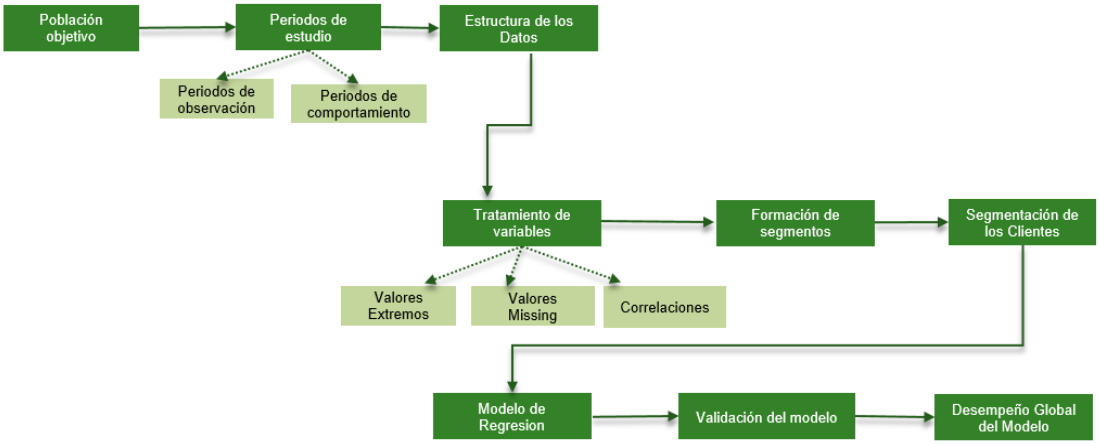
Modificación	Realizado por	Fecha de Realización	Fecha de Aprobación
Formatos	Carmen Ñique Chacón	27/08/2019	

1. INTRODUCCIÓN

Las entidades financieras están en constante cambio sobre el comportamiento tradicional de acercamiento con sus clientes, a una conducta mucho más proactiva que consiga adaptar su cartera de productos y servicios a las necesidades particulares de cada persona.

En este documento pretendemos estimar el ingreso que pueda tener un cliente determinado, en función a la información que se tenga sobre él, con el objetivo de poder ofrecer el producto o servicio que más se acomode a lo que podría necesitar el cliente. Para la estimación del ingreso de clientes se ha ido desarrollando una serie de modelos de regresión múltiple, así como técnicas de segmentación de clientes.

Gráfico 1 Flujo metodológico para la construcción del modelo



2. FUENTES DE INFORMACIÓN

Las fuentes de información que se consideraron para la construcción del modelo de ingreso corresponden a información del Banco Falabella Perú, información del sistema financiero y sociodemográfica.

En la información interna del banco tenemos:

- Pagos: información de los abonos que realiza el cliente para amortizar su deuda en el Banco Falabella Perú.
- Banco Falabella: Son los montos desembolsados de los productos Rapicash o Supercash.
- Transaccional: Son los montos de compras realizadas en los distintos retail del grupo (Tottus, Saga Falabella, Sodimac y Maestro).

En la información del sistema Financiero tenemos:

- Los montos de saldos por distintos tipos de producto (hipotecario, tarjeta de crédito, microempresa, etc.).
- Montos de Líneas
- Número de créditos.

También contamos con la información de la Superintendencia Nacional de Registros Públicos (Sunarp) de la cual se obtiene un *flag* de la tenencia de vehículos.

Finalmente, la información de ingreso es obtenidas de cuentas sueldo del Banco Falabella consolidadas a diciembre de 2018.

Tabla 1 Variables según fuente de información

Sociodemográfica	Transaccional	Banco Falabella	Pagos en el Banco Falabella	Sistema Financiero (RCC)
•Sexo •Estado Civil •Siituación Laboral •Marca de Vehiculo •Ubicacion •Edad •Grado de Instrucción.	•Número de Transacciones en Maestro. •Monto Ticket en Maestro. •Monto Ticket en Sodimac. •Monto Ticket en Tottus. •Monto Ticket en Saga Falabella. •Consumo en Restaurantes. •Consumo en Educación. •Gasto en Salud. •Gasto en entretenimiento.	•Desembolso RapiCash. •Desembolso Supercash. •Linea TC. •Saldo TC.	•Pagos realizados por los clientes. •Pago correspondiente al mes según Estado de Cuenta. •Pago Mínimo del estado de cuenta.	•Meses Normal. •Cantidad de entidades que reporta. •Número de prestamos. •Saldo total en el sistema financiero. •Saldo credito comercial. •Saldo Microempresa. •Saldo TC Compras. •Saldo Dispocisión de efectivo. •Linea TC en el sistema Financiero. •Saldo Credito Vehicular. •Saldo Credito Hipotecario.

Se adjunta el archivo donde se encuentra la descripción de las variables de las fuentes de información:

Fuentes de Información.xlsx

3. POBLACIÓN OBJETIVO

La población objetivo para el modelo de ingresos de clientes son todos los clientes activos del Banco Falabella que cuentan con una tarjeta hasta diciembre 2018 (Stock), siendo excluidos aquellos que se encuentren en la base de personas fraudes durante todo el año 2018.

4. PERIODOS DE ESTUDIO

El periodo de estudio comprende los periodos de observación y periodos de comportamiento.

El periodo de observación hace referencia al periodo en donde la población objetivo es seleccionada para su evaluación. Mientras que los periodos de comportamiento se refieren a los periodos en los cuales se recogerán la información histórica de los clientes.

4.1. PERIODO DE OBSERVACIÓN

El periodo de observación para la estimación del ingreso es diciembre 2018, en donde se evaluarán a todos los clientes de la población objetivo.

Stock de clientes Banco Falabella diciembre 2018

4.2. PERIODO DE COMPORTAMIENTO

Los periodos de comportamiento que se han considerado, para cada fuente de información, son los siguientes:

Fuente 1.

- Compras en los Retail del grupo: es decir las compras generadas por los clientes en TOTTUS, SAGA FALABELLA, SODIMAC, MAESTRO y CONVENIOS.

Tabla 2 Distribución de montos de transacción por Periodo, según Retail

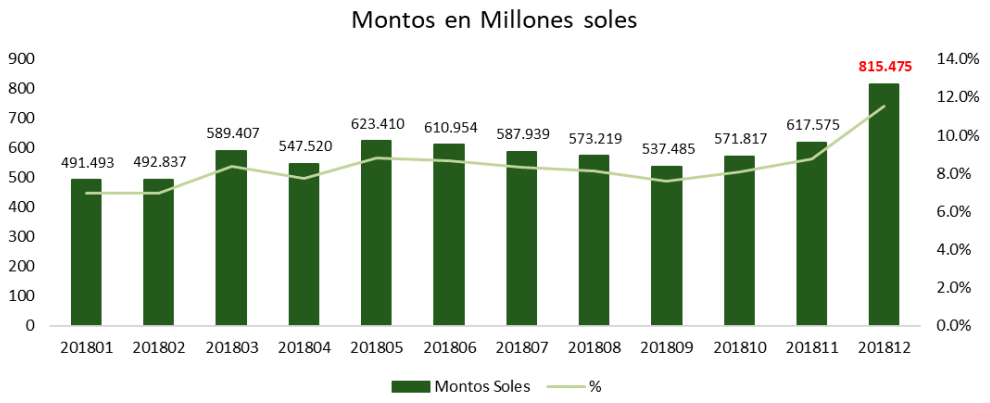
	Periodos											
	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn
	201801	201802	201803	201804	201805	201806	201807	201808	201809	201810	201811	201812
TOTTUS	129,883,486	112,825,796	131,289,575	134,064,268	169,099,296	145,037,836	134,790,503	144,734,216	129,457,240	144,231,073	152,713,510	211,398,116
FALABELLA	140,980,441	125,586,183	157,034,952	162,890,871	207,338,662	215,910,913	187,519,176	146,200,302	132,213,561	120,069,410	136,176,631	247,565,708
MAESTRO	52,168,160	47,251,009	51,559,915	48,497,652	47,837,543	48,203,520	54,392,749	55,539,253	48,651,933	49,414,190	52,376,891	55,904,736
SODIMAC	83,307,981	77,023,121	81,533,223	78,453,180	78,910,452	80,578,953	91,155,844	90,014,756	80,297,132	76,970,917	82,358,427	94,215,786

- Convenios*:* Aquí se encuentran los consumos realizados con la tarjeta en distintos comercios ajenos a los Retail del grupo, siendo excluidas aquellas transacciones del mes de diciembre, ya que se presentan picos altos en el total de montos transaccionados (815 millones de soles).

Tabla 3 Distribución de montos de convenios por periodo

Periodos											
Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn	Montos Trxn
201801	201802	201803	201804	201805	201806	201807	201808	201809	201810	201811	201812
491,492,745	492,837,050	589,406,800	547,519,858	623,410,305	610,953,540	587,938,539	573,218,792	537,484,917	571,816,789	617,574,706	815,475,435

Gráfico 2 Distribución de montos de Convenios por periodo



Fuente 2.

- Pagos: Hace referencia a los pagos que el cliente realiza a su tarjeta de crédito a cierre de mes; siendo utilizados como variables predictoras del modelo de ingreso.

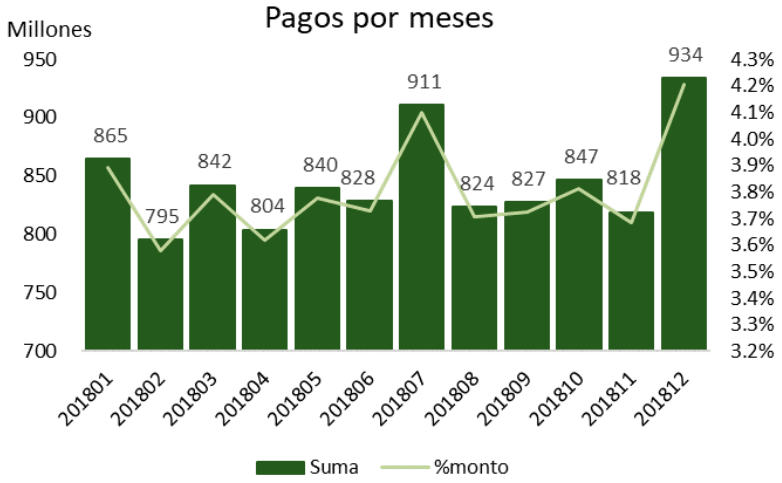
Tabla 4 Distribución de pagos por periodo

Periodos											
Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos	Montos pagos
201801	201802	201803	201804	201805	201806	201807	201808	201809	201810	201811	201812
821,014	813,270	840,973	834,703	871,204	865,055		881,005	892,023	909,039	947,152	

*: Número de clientes con montos mayor a cero

MESES QUE SE EXCLUIRÁN

Gráfico 3 Distribución de montos de Pagos por periodo



Como se puede observar en la gráfica se presentan picos altos en los periodos de julio y diciembre, e cual es explicado por el pago doble que existe en estos meses, lo cual es un comportamiento atípico según la serie.

Fuente 3

- Sistema Financiero (RCC): La temporalidad considerada para la información en el sistema financiero es de 201710 al 201809

Tabla 5 Total de Registros por periodo

	Periodos											
	201710	201711	201712	201801	201802	201803	201804	201805	201806	201807	201808	201809
n° registros	9,915,020	9,968,758	9,999,815	10,050,484	10,243,150	10,104,935	10,128,852	10,156,774	10,206,227	10,100,797	10,264,088	10,310,115

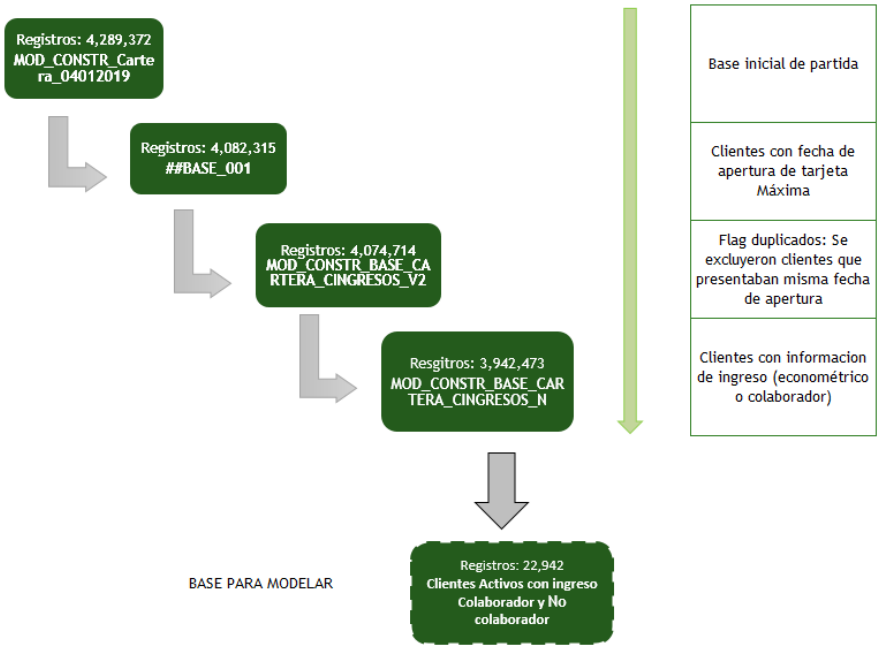
5. ANÁLISIS INICIAL DE DATOS

Las variables que se obtuvieron de las distintas fuentes de información se dividen en: Variable Ingresos, Variables Numéricas y Variables Categóricas.

5.1. VARIABLE INGRESOS

Para la construcción de la base de ingresos con el cual se modelará, se sigue el siguiente flujo:

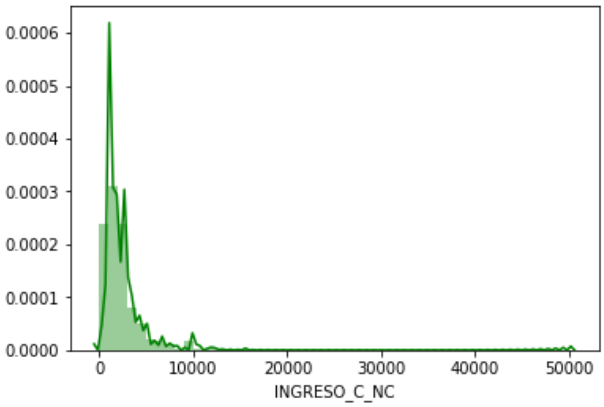
Tabla 6 Flujo construcción base para modelar



Iniciamos con la base de clientes de diciembre 2018 (4,289,372 clientes) y se realizan las depuraciones correspondientes, quedándonos con los clientes activos y aquellos que no se encuentren en la base de Fraudes de todo el 2018 (22 mil 942 registros).

n	22,942
Promedio	2,465
Min.	0.01
Max.	50,000
Percentil 10	970
Percentil 20	990
Percentil 30	1,200
Percentil 40	1,507
Percentil 50	1,798
Percentil 60	2,397
Percentil 70	2,548
Percentil 80	3,160
Percentil 90	4,700

Gráfico 4 Distribución de Ingresos Cuenta Sueldo de Clientes Activos



5.2. VARIABLES CONTINUAS

Se procedió a obtener una matriz de información con estadísticas descriptivas básicas como, por ejemplo: Mínimo, Máximo, Promedio, Percentiles, % de outliers, etc.

Analisis_univariado.xlsx

Analizando el porcentaje de missing como una primera medida de depuración de variables, obtendremos una reducción en la cantidad de variables numéricas.

El porcentaje de missing tolerables fue establecido de a lo más 15% de registros.

Acciones Variables Numericas.xlsx

5.3. VARIABLES NOMINALES

En el siguiente cuadro se tienen las variables nominales que se obtuvieron de la fuente sociodemográfica. Cabe mencionar, que en cada cuadro se muestra la cantidad de registros missing o vacíos que más adelante se explica la metodología de imputación para estos casos.

Tabla 7 Variables Discretas

Sexo	n
F	674,170
M	598,206
X	11
missing	403

EstadoCivil	n
Casado	216,959
Divorci	515
Separad	2,153
Soltero	1,046,829
Viudo	5,925
	6
missing	403

NSE	n
A	66,816
B	561,812
C	302,378
D	15,125
E	3,508
missing	323,151

Situacion Laboral	n
DEP	625,896
INDEP	417,019
INFOR	229,472
missing	403

Ubicación	n
CALLAO	71,125
LIMA CENTRO	266,757
LIMA ESTE	143,514
LIMA NORTE	160,348
LIMA SUR	125,392
ZONA CENTRO	146,826
ZONA NORTE	239,324
ZONA SUR	111,992
no se encuentra	7,109
missing	403

Grupo Grado Instrucción	n
0.sin categoria	103,430
1.Grupo 01	1,433
2.Grupo 02	5,927
3.Grupo 03	109,312
4.Grupo 04	674,068
5.Grupo 05	97,428
6.Grupo 06	281,192

6. METODOLOGÍA DEL TRATAMIENTO DE VARIABLE

6.1. CONSTRUCCIÓN DE VARIABLES

Para la construcción de variables de las distintas fuentes de información se consideraron el sentido de la variable en el negocio, su facilidad en cálculo e interpretación.

En el siguiente archivo se presentan las variables construidas, por fuente de información.

Construcción_Variables.xlsx

6.2. TRATAMIENTO DE OUTLIERS

Para el tratamiento de outliers se utilizó la metodología de 95% de la concentración de datos. Se identificó el quantil 0.025, y aquellos valores que se encuentren por debajo del mismo serán acotados con dicho valor. Los valores que se encuentren por encima del quantil 0.975 serán reemplazados por dicho valor.

Se aplicó este criterio aquellas variables que contaban a lo más el 5% de datos outliers

6.3. TRATAMIENTO DE MISSING

El tratamiento que se dio para los valores vacíos, en este caso para las variables numéricas fue reemplazar los vacíos por **"0"**, esto dado que realmente estos vacíos corresponden a valores nulos, por ejemplo: Un valor vacío en una variable pagos corresponde a un pago igual a cero, al igual que un monto vacío de transacción en el retail corresponde a un monto cero (representa la no transacción de un cliente).

Por otro lado, en el caso de las variables categóricas, se ha optado por el criterio de imputación respetando las proporciones que guarda cada una de las categorías.

Esto es, por ejemplo:

Se tiene la variable "Situación Laboral"

Situación Laboral	n
DEP	625,896
INDEP	417,019
INFOR	229,472
missing	403

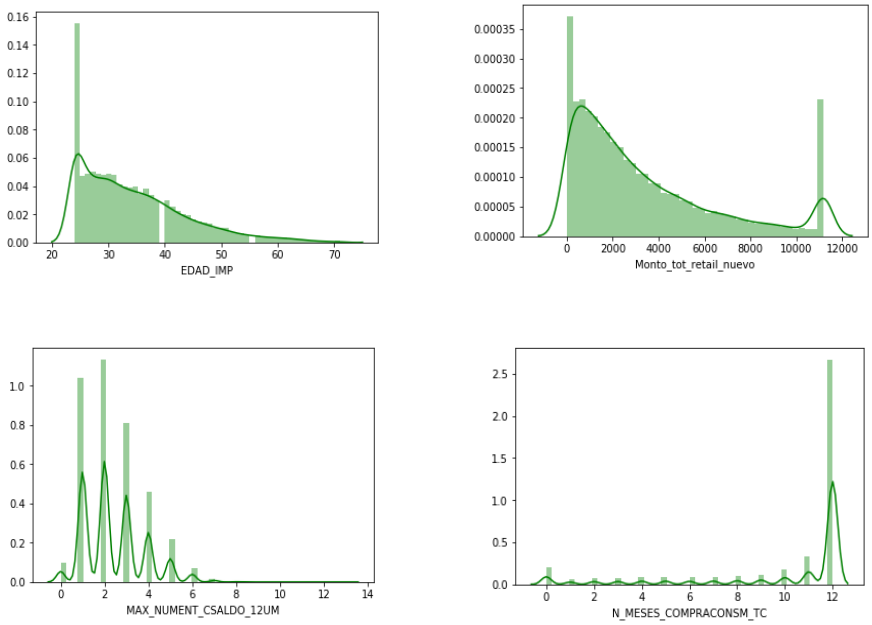
Las proporciones para los 403 missing

Situación Laboral	%	Proporción
DEP	49.18	0.49
INDEP	32.76	0.33
INFOR	18.03	0.18

6.4. ANÁLISIS UNIVARIANTE

Luego de hacer los tratamientos de outliers y missing a las variables, se obtiene la distribución de algunas de las variables como se muestra.

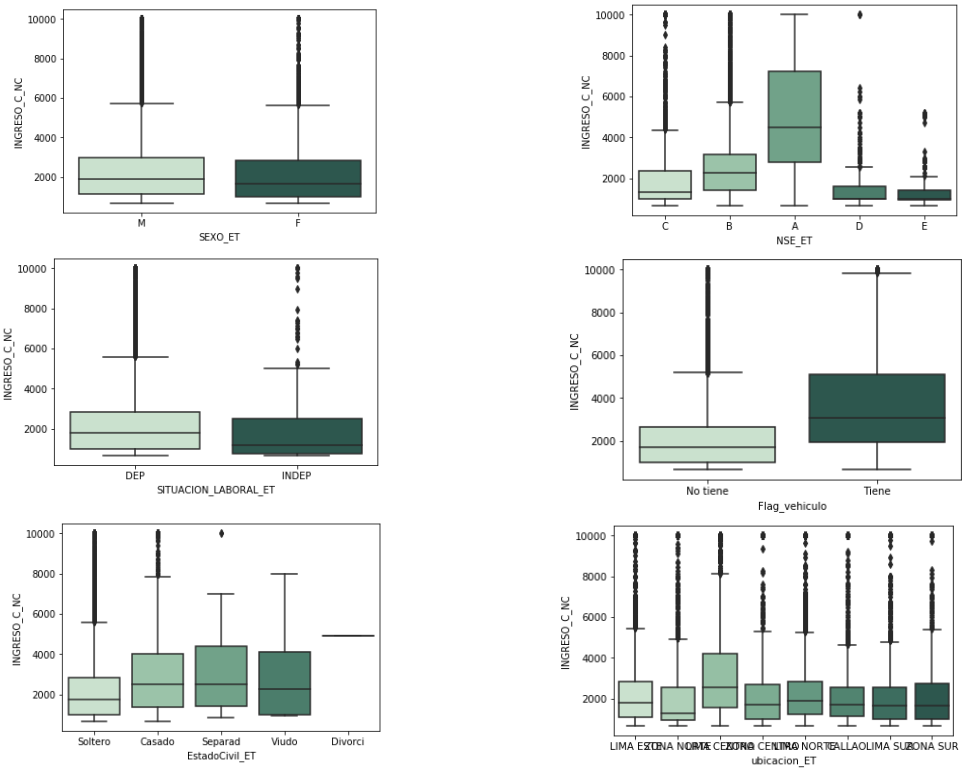
Gráfico 5 Distribuciones por variables



6.5. ANÁLISIS BIVARIANTE

Para el análisis bivariante se tienen las siguientes gráficas de cajas del ingreso vs. Las distintas variables categóricas.

Gráfico 6 Distribución del ingreso según variables



6.6. ANÁLISIS DE CORRELACIONES

Antes de hacer los análisis respectivos de las correlaciones, fue necesario retirar las variables referentes al saldo en el sistema financiero y líneas en el banco, dado que por recomendaciones del negocio no debían ser considerados para la estimación de ingresos.

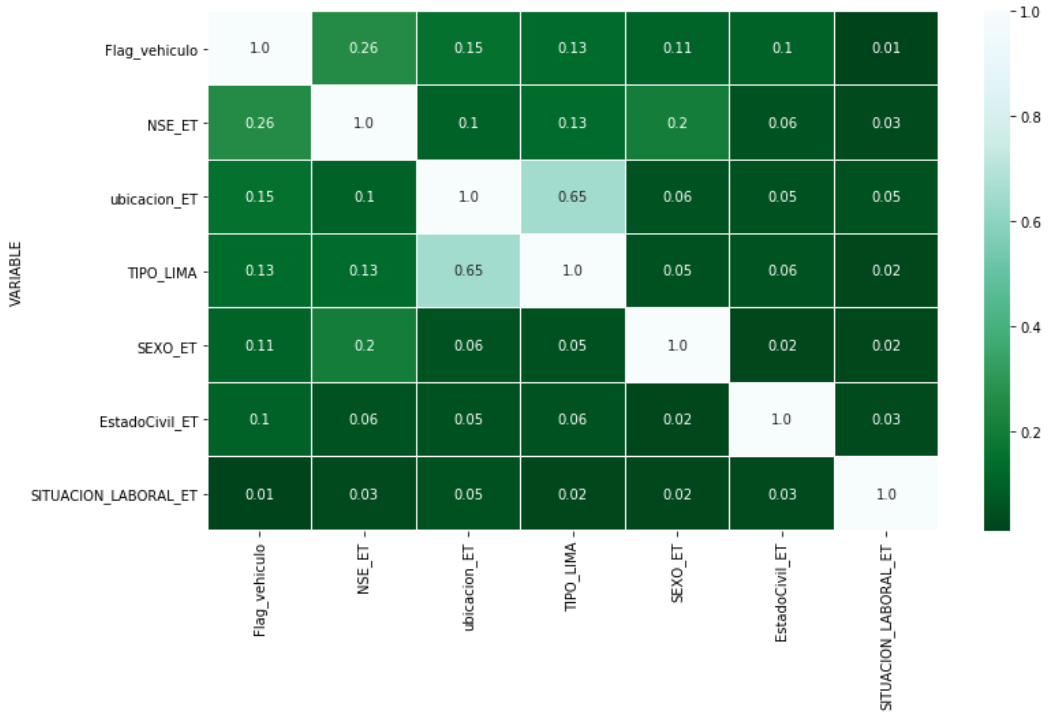
El análisis de las correlaciones entre las variables que son candidatas para el modelo es de mucha importancia para que no exista problema en los ajustes.

Correlación de Pearson: Tenemos la correlación de las variables continuas para los 22,942 registros con el método de Pearson, las cuales se encuentran adjuntas en el siguiente archivo.

Matriz_Correlacion_modelar.xlsx

Se tiene que la variable de mayor correlación con el ingreso es el saldo total en el sistema financiero (0.38), el cual me da una ligera idea de que ésta puede ser una variable predictora en el modelo. También se tienen correlaciones negativas con la variable que hace referencia a la utilización de la tarjeta.

Correlación V de Cramer: Por otro lado, se calcula la correlación de las variables cualitativas con el método de V de Cramer.



De la cual tenemos que si los valores de V Cramer se encuentran cercanos a 0 significa que no existe correlación entre las variables categóricas, sin embargo, si se encuentra cercano a 0.6 existe una correlación relativamente intensa.

7. CONSTRUCCIÓN DEL MODELO

7.1. SEGMENTACIÓN (CLÚSTER)

Para realizar una segmentación de clientes, es necesario tener en claro cómo deseo obtener la segmentación, esto es tener definido en función de que variable o variables deseo hacer la segmentación de clientes.

En un modelo de ingresos la variable clave para hacer la segmentación tiene que ser el ingreso que percibe el cliente, para el cual es necesario realizar un análisis de agrupación de los ingresos.

El análisis de clúster es un método que permite descubrir asociaciones y estructuras en los datos que no se tiene a priori pero que pueden ser útiles una vez que se encuentren.

Clúster Jerárquico Aglomerativo

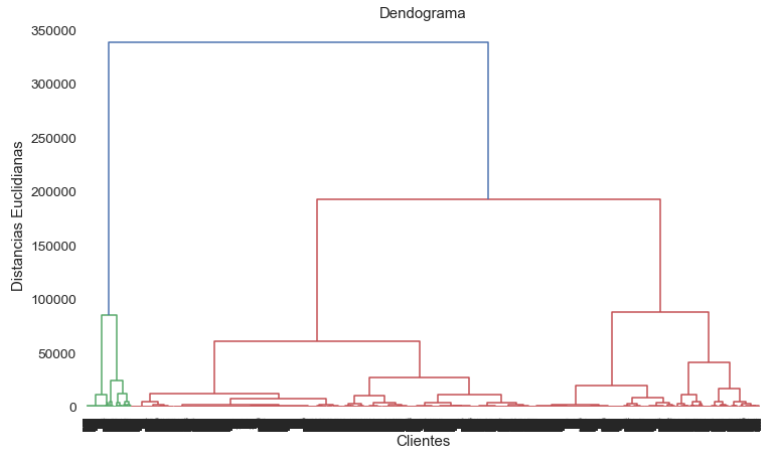
Los algoritmos de agrupación jerárquica aglomerativo son aquellos que parten de una fragmentación completa de los datos, estos se van fusionando hasta conseguir una situación contraria, es decir, todos los datos se unen en un solo grupo.

Dendrogramas

El dendrograma es un diagrama que muestra las agrupaciones sucesivas que genera un algoritmo jerárquico aglomerativo.

Para la base de clientes (22942 clientes) se realizó la gráfica del dendograma, como se presenta:

Gráfico 7 Dendograma de Ingresos



Como se puede observar en el dendograma, tenemos la ligera idea de formar 3 grupos, por lo que a continuación se muestran las estadísticas básicas al formar estos 3 clúster:

Tabla 8 Frecuencias por Clúster

Clusters	n	%
Clúster 0 (medios)	7653	33.36%
Clúster 1 (altos)	1540	6.71%
Clúster 2 (bajos)	13749	59.93%
TOTAL	**22942**	**100.00%**

Estadísticas del ingreso por cada Clúster generado.

Tabla 9 Estadísticas por Clúster

Estadísticas	Clúster 0	Clúster 1	Clúster 2
Min.	2,392	5,637	671
Max.	5,621	10,000	2,390
Media	3,259	8,327	1,316
Percentil 10	2,500	6,133	904
Percentil 20	2,548	6,622	970
Percentil 30	2,548	6,859	990
Percentil 40	2,850	7,600	1,002
Percentil 50	2,993	8,000	1,198
Percentil 60	3,150	10,000	1,385
Percentil 70	3,500	10,000	1,600
Percentil 80	4,000	10,000	1,700
Percentil 90	4,700	10,000	1,943

Como es de notarse, tenemos un grupo muy pequeño (6.71% ingresos altos), el cual nos puede generar problemas para predecir más adelante.

Por lo tanto, se decidió reagrupar el Clúster 0 y el Clúster 1 (Ingresos medios- Altos), teniéndose las siguientes distribuciones y estadísticas.

Tabla 10 Frecuencias por Clúster Reagrupados

Clusters	n	%
Clúster 1 (Medio-Alto)	9,193	40.1%
Clúster 2 (Bajos)	13,749	59.9%
TOTAL	**22942**	**100.0%**

Gráfico 8 Distribución de Ingresos por Clúster Reagrupado

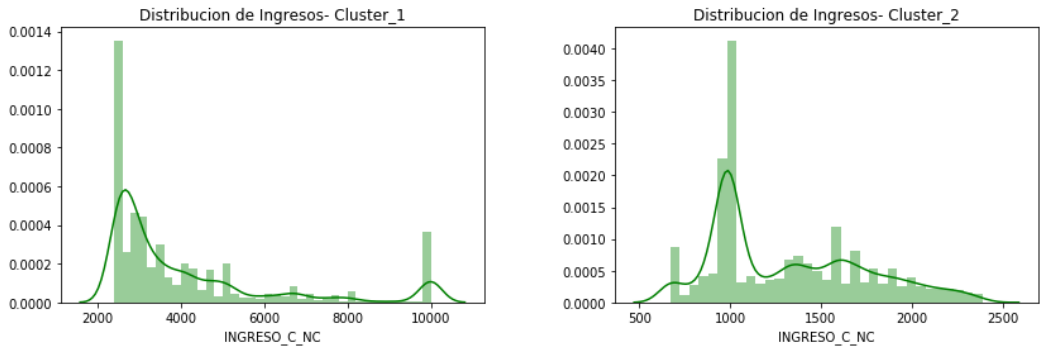


Tabla 11 Estadísticas por Clúster Reagrupado

Resumen	Clúster 1 (I. Medios- Altos)	Clúster 2 (I. Bajos)
Min.	2,392	671
Max.	10,000	2,390
Media	4,108	1,316
Percentil 10	2,548	904
Percentil 20	2,548	970
Percentil 30	2,700	990
Percentil 40	2,895	1,002
Percentil 50	3,157	1,198
Percentil 60	3,568	1,385
Percentil 70	4,200	1,600
Percentil 80	5,077	1,700
Percentil 90	7,600	1,943

7.2. MODELO ÁRBOL DE DECISIONES

Los árboles de inferencia condicional estiman una relación de regresión mediante la partición recursiva binaria en un marco de inferencia condicional.

El algoritmo funciona de la siguiente manera[1]:

1. Pruebe la hipótesis nula global de independencia entre cualquiera de las variables de entrada y la respuesta (que también puede ser multivariable). Detente si esta hipótesis no puede ser rechazada. De lo contrario, seleccione la variable de entrada con la asociación más fuerte a la respuesta. Esta asociación se mide por un valor de p correspondiente a una prueba para la hipótesis nula parcial de una sola variable de entrada y la respuesta.
2. Implementar una división binaria en la variable de entrada seleccionada.
3. Repite repetidamente los pasos 1) y 2).

El modelo de segmentación que se realizó tuvo como variables input:

Pago Promedio en los 6 últimos meses.

Utilización promedio en los últimos 6 meses.

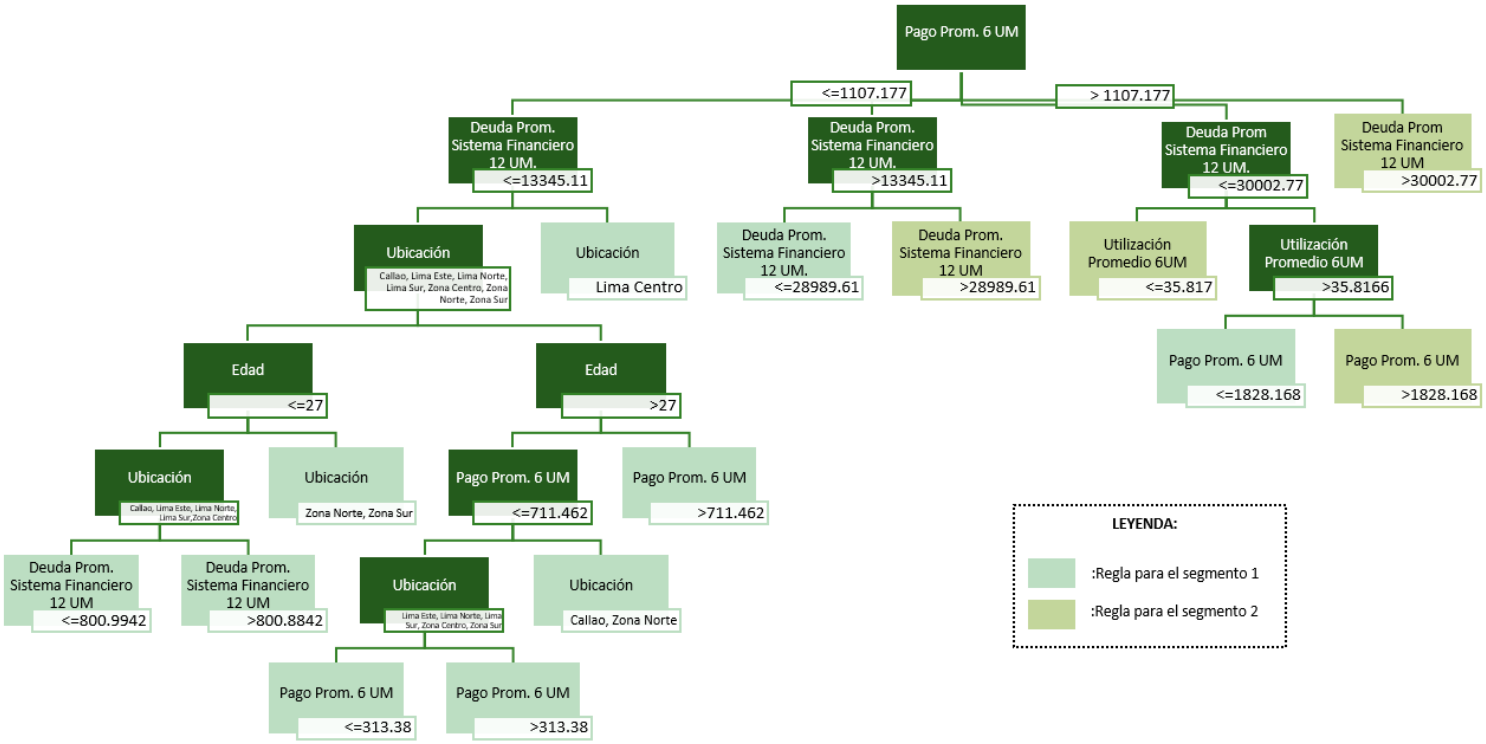
La deuda promedio en los 12 últimos meses del sistema financiero.

La edad.

La ubicación según Reniec

[1] (Hothorn, Hornik , & Zeileis, 2006)

Gráfico 9 Árbol de decisiones



En la construcción del modelo de segmentación se tienen los siguientes indicadores:

La matriz de Confusión que se obtuvo es la siguiente

Tabla 12 Matriz de confusión

		TRAIN		TEST	
		Real		Real	
		I. Medios- Altos	I. Bajos	I. Medios- Altos	I. Bajos
Predicción	I. Medios- Altos	67.05	32.95	67.28	32.72
	I. Bajos	29.26	70.74	29.69	70.31

Finalmente se tiene una precisión del 70% aproximadamente, lo cual india que nuestro modelo predice correctamente al 70% de los clientes.

Tabla 13 Precisión del Modelo

	TRAIN	TEST
Predicción	69.69	69.45

7.3. MODELOS DE REGRESIÓN ROBUSTA

La regresión robusta es un método de regresión que se usa cuando la distribución del residual no es normal o hay algunos valores atípicos que afectan el modelo.[2]

Si asumimos que nuestros datos siguen un modelo de regresión lineal normal, las estimaciones y las pruebas de mínimos cuadrados funcionan bastante bien, pero no son sólidas cuando el supuesto de normalidad para la población de errores aleatorios no es válido. La regresión M se desarrolló específicamente para ser robusta con respecto a este supuesto. Peter Huber introdujo la idea de la estimación M en 1964.

Huber (1964) construyó su estimación M para que fuera óptima si se supone que la distribución del error es una distribución normal contaminada por una pequeña proporción de errores de alguna distribución arbitraria.

- **Modelos**

El modelo de árbol de decisiones etiqueto a 16422 clientes en el segmento 1, de los cuales se retiraron aquellos clientes que no debieron ser clasificados en este segmento (ingresos que se encuentran por debajo de 2390), quedándonos con 11602 registros.

Por otro lado, el modelo clasificación a 6520 clientes en el segmento 2 de los cuales se retiraron los clientes con ingresos por debajo de 2390, clientes mal clasificados, obteniendo 4374 clientes en el segmento 2.

- **Selección de Variable**

Para la elección de las variables predictoras del ingreso en este segmento de ingresos bajos, se utilizó la metodología de selección de variables Forward.

El método Forward comienza con el modelo nulo, sin ningún predictor para ir incluyéndolos de uno en uno hasta incluirlos todos. En cada paso la variable que aporta mayor mejora es añadida al modelo. Luego de generar los modelos para la estimación de ingresos, procedemos a comprarlos entre ellos con el Criterio de Información Bayesiano (BIC) la cual es una medida de bondad de ajuste de un modelo estadístico, y es a menudo utilizado como un criterio para para la selección de modelos entre un conjunto finito de modelos.

En los archivos adjuntos se encuentran las iteraciones realizadas.

Seleccion_Variables_Solo_Segmento1.xlsx

Seleccion_Variables_Solo_Segmento2.xlsx

Sin embargo, el obtener un conjunto de variables proporcionadas por la metodología Forward, se adicionó a ello los criterios del negocio y la correlaciones.

(Hasih Pratiw & Twenty Liana4, 2014)[2]

- **Muestra**

El criterio para dividir las bases en train (construcción) y test (validación) que se consideró fue, la muestra de train del modelo debe de ser del 80%, es decir se construye el modelo con el 80% de los registros. Estos registros fueron tomados de manera aleatoria y considerando una semilla en cada segmento, con el propósito de poder obtener la misma muestra si se quisiera volver a replicar.

Mientras que para la validación se tomó el 20% restante en cada segmento.

- **Variables de los modelos**

La forma de cálculo de cada una de las variables se encuentra en el ANEXO II.

En el Segmento 1 el modelo cuenta con 7 variables y una constante

Tabla 14 Variables del Modelo Segmento 1

VARIABLES	DESCRIPCIÓN
UTIL_MAXIMO_6UM	Utilización de línea TC máxima en los 6 meses
UBICACION_ET_ZONA NORTE	Pertenencia a la Zona Norte del Perú
N_MESES_LINEA_TC	Meses con línea de TC en los 12 meses
PAGO_TOTAL_6UM	Pago total en los últimos 6 meses
EDAD_IMP	Edad
MAX_NUMTC_CSALDO_12UM	Máximo número de tarjeta de crédito con saldo en 12 meses
UBICACION_ET_ZONA SUR	Pertenencia a la Zona Sur del Perú

En el Segmento 2 el modelo cuenta con 6 variables y una constante

Tabla 15 Variables del Modelo Segmento 2

VARIABLES	DESCRIPCIÓN
N_MESES_DISEF_TC	Meses con saldo disposición de efectivo en los 12 meses
PAGO_MEDIANA_6UM	Pago mediana en los últimos 6 meses
FLAG_VEHICULO_DUMMY	Tenencia de Vehículo
UTIL_PROM_6UM	Utilización de línea TC (Línea consumo + línea rapicash) promedio en los 6 últimos meses
UBICACION_ET_LIMA CENTRO	Pertenencia a la Lima Centro
MONTO_TOT_RETAIL	Monto total de transacción en los retail del grupo Falabella (Sodimac, Maestro, Saga Falabella y Tottus)

Los parámetros estimados por el modelo de Regresión Huber son los siguientes

Tabla 16 Parámetros Modelo Segmento 1

Variables	Beta	Desv. Estándar Error	P-Valor	Intervalo de Confianza [0.025	0.975]
Constante	1167.1037	21.846	0	1124.286	1209.921
UTIL_MAXIMO_6UM	-1.9165	0.123	0	-2.158	-1.675
UBICACION_ET_ZONANORTE	-147.6182	10.252	0	-167.172	-127.524
N_MESES_LINEA_TC	4.4156	1.405	0	1.662	7.169
PAGO_TOTAL_6UM	0.0259	0.002	0	0.022	0.03
EDAD_IMP	2.5977	0.483	0	1.651	3.544
MAX_NUMTC_CSALDO_12UM	36.0056	4.195	0	27.784	44.227
UBICACION_ET_ZONASUR	-82.665	18.935	0	-119.777	-45.553

Ecuación del modelo:

$$Y=1167.1037-1.9165*x_1-147.6182*x_2+4.4156*x_3+0.0259*x_4+2.5977*x_5+36.2256*x_6-82.665*x_7$$

X_1 : Utilización de linea TC máxima en los 6 ultimos meses

X_2 : Pertenece a Zona Norte del Perú

X_3 : Meses con linea de TC en los 12 meses

X_4 : Pago total en los últimos 6 meses

X_5 : Edad

X_6 : Máximo número de tarjeta de crédito con saldo en 12 meses

X_7 : Pertenencia a la Zona Sur del Perú

Tabla 17 Parámetros Modelo Segmento 2

Variables	Beta	Desv. Estándar Error	P-Valor	Intervalo de Confianza	
				[0.025	0.975]
Constante	3688.0678	83.881	0	3523.665	3852.471
N_MESES_DISEF_TC	-54.6302	7.235	0	-68.81	-40.451
PAGO_MEDIANA_6UM	0.1425	0.041	0.001	0.061	0.224
FLAG_VEHICULO_DUMMY	907.3741	69.814	0	770.541	1044.207
UTIL_PROM_6UM	-5.3424	1.05	0	-7.401	-3.284
UBICACION_ET_LIMACENTRO	686.4273	67.405	0	554.316	818.538
MONTO_TOT_RETAIL	0.0532	0.009	0	0.035	0.072

Ecuación del modelo:

$$Y=3688.0678-54.6302*x_1+0.1425*x_2+907.3741*x_3-5.3424*x_4+686.4273*x_5+0.0532*x_6$$

X_1 : Meses con saldo disposicion de efectivo en los 12 meses

X_2 : Pago mediana en los ultimos 6 meses

X_3 : Tenencia de Vehículo

X_4 : Utilización de Linea TC promedio en los 6 ultimos meses

X_5 : Pertenencia a Lima Centro

X_6 : Monto total de transaccionen los retail del grupo Falabella (Tottus,Sodimac,Maestro y Saga Falabella)

Como infomación resumen de los modelos de regresion ver el ANEXO III

- **Supuestos del Modelo**

a) Homocedasticidad (contraste de Goldfeld-Quandt.[3])

Parte del supuesto de que la varianza de la perturbación, σ_i^2 depende monótonamente de los valores de una variable Z_i , que puede ser o no uno de los regresores del modelo. En cualquier caso, debe ser una variable observable. Para contrastar la hipótesis nula de ausencia de heterocedasticidad

$$H_0:\sigma_i^2=\sigma_u^2 \text{ (homocedasticidad)}$$

contra la alternativa de existencia de heterocedasticidad:

$$H_0:\sigma_i^2=\sigma_u^2*Z_i \text{ para algún i=2,…,p (heterocedasticidad)}$$

b) Independencia (Estadístico Durbin Watson)

Independencia entre los residuos mediante el estadístico de Durbin-Watson que toma valor 2 cuando los residuos son completamente independientes (entre 1.5 y 2.5 se considera que existe independencia), DW2 autocorrelación negativa

$$DW=\frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \quad ; \quad 0 \le DW \le 4$$

c) Multicolinealidad

La multicolinealidad en regresión es una condición que ocurre cuando algunas variables predictoras que se incluyen en el modelo están correlacionadas con otras variables predictoras del mismo.

Los resultados de la prueba de los supuestos en cada segmento se encuentran en el archivo adjunto.

Modelos
Regresion_Supuestos.

- **Factor de Ajuste:**

Para poder garantizar que la sobrestimación de los ingresos se encuentre dentro de lo establecido por la SBS (sobrestimación<25%) se proponen factores de ajustes que se aplicarán a aquellos ingresos estimados mayores a 1000. Se realizó una serie de iteraciones en Python (ANEXO II) para poder encontrar el mejor ajuste que nos proporcione porcentajes que se encuentren dentro de lo permisible. La forma del factor a aplicarse fue:

- Dar un peso a la estimación (Reduce la estimación)
- Conservar la forma de la distribución haciendo un desplazamiento de la misma con ayuda de la desviación estándar.

[3] (Gallegos Gómez, 2008-2009)

El factor escogido es:

Ingreso Estimado*0.90-0.15*Desviación Estándar

8. DESEMPEÑO DEL MODELO

En la estimación de ingresos los indicadores de desempeño del modelo son:

- ✓ La Subestimación que está definida como:

$$\frac{(\text{Ingreso Estimado-Ingreso Real})}{\text{Ingreso Real}} \leq -0.25$$

- ✓ Correctos definida como:

$$-0.25 < \frac{(\text{Ingreso Estimado-Ingreso Real})}{\text{Ingreso Real}} \leq 0.25$$

- ✓ La sobreestimación definida como:

$$0.25 < \frac{(\text{Ingreso Estimado-Ingreso Real})}{\text{Ingreso Real}}$$

También se considera el error relativo promedio como un indicador para el desempeño de un modelo.

En el siguiente cuadro se muestran los rangos de variaciones del Ingreso Estimado respecto al ingreso real.

Tabla 18 Variaciones de los Ingresos

Rango Variacion	n	%	%acum
01.<=-0.50	665	4.2%	100.0%
02.<-0.50 a -0.25]	3980	24.9%	95.8%
03.<-0.25 a 0>	4238	26.5%	70.9%
05.<0 a 0.25]	4827	30.2%	44.4%
06.<0.25 a 0.50]	1511	9.5%	14.2%
07.<0.50 a 0.75]	589	3.7%	4.7%
08.<0.75 a 0.90]	129	0.8%	1.0%
09.<0.9-1.00]	28	0.2%	0.2%
10.>1.00	10	0.1%	0.1%

Como indicadores globales se tiene que el modelo de ingresos sobreestima el 14.2% del total de registros.

Subestimación	29.1%
Correctos	56.7%
Sobreestimación	14.2%

El error relativo del modelo es:

Error Relativo	0.24

ANEXOS

ANEXO I

```
1.  ############### MODELO DE CLASIFICACION ########################
2.  ### Cargando la Base
3.  BASE_CLUSTER_AGLOM_3_07_2019_V2 <- read.csv("BASE_CLUSTER_AGLOM_20_08_2019.csv
    ", header = T,sep = "|")
4.
5.  ### Filtrando las variables que entraran al modelo
6.
7.  BASE_CLUSTER_SELEC=BASE_CLUSTER_AGLOM_3_07_2019_V2[,c("Cluster_Aglom_FINAL",

8.                                                      "ubicacion_ET",
9.                                                      "Flag_vehiculo",
10.                                                     "UTIL_PROM_6UM",
11.                                                     "PROM_SLDTOT_SF_12UM",
12.                                                     "PAGO_PROMEDIO_6UM",
13.                                                     "EDAD_IMP")]
14.
15. # Frecuencia de los segmentos
16. table(BASE_CLUSTER_SELEC$Cluster_Aglom_FINAL)
17.
18.
19. # Convirtiendo a factor las variables categoricas
20.
21. BASE_CLUSTER_SELEC$Cluster_Aglom_FINAL<- as.factor(BASE_CLUSTER_SELEC$Cluster_
    Aglom_FINAL)
22. BASE_CLUSTER_SELEC$ubicacion_ET<- as.factor(BASE_CLUSTER_SELEC$ubicacion_ET)
23. BASE_CLUSTER_SELEC$Flag_vehiculo<- as.factor(BASE_CLUSTER_SELEC$Flag_vehiculo)

24.
25. # Arbol de Clasificación para los segmentos
26.
27. # Libreria
28.
29. library(party)
30.
31.
32. #### Dividiendo la base en train y test
33.
34. set.seed(21)  ## semilla
35. sample_ind <- sample(nrow(BASE_CLUSTER_SELEC),nrow(BASE_CLUSTER_SELEC)*0.80)
36. train <- BASE_CLUSTER_SELEC[sample_ind,]
37. test <- BASE_CLUSTER_SELEC[-sample_ind,]
38.
39. ## Arbol de Decisiones
40. ARBOL_CLUST <- ctree(train$Cluster_Aglom_FINAL ~ . , data = train, controls =
    ctree_control(minbucket = 1000,
41.
                  mincriterion = 0.95, minsplit =500))
42.
43.
44. # Graficando el arbol
45. plot(ARBOL_CLUST)
46.
47. # Imprimiendo las reglas
48. print(ARBOL_CLUST)
49.
50. ###### TRAIN ######
51. ## Matriz de confusion- train
52.
53. #agregando los nodos finales
54. train$prediccion <- predict(ARBOL_CLUST, train, type="response")
55.
56. matriz_conf=table(train$prediccion, train$Cluster_Aglom_FINAL)
```

```
prop.table(matriz_conf, 1) * 100

## Presición del modelo - train
print(100 * sum(diag(matriz_conf)) / sum(matriz_conf))

## NODOS POR SEMENTO
train$NODO<-predict(ARBOL_CLUST, type = "node",train)
#### # Registros en los nodos por segmento
tab <- table(predict(ARBOL_CLUST, type = "node",train), train$Cluster_Aglom_FINAL)
prop.table(tab, 1) * 100


######## TEST ###########

# Predicciones
test$prediccion <- predict(ARBOL_CLUST, test, type="response")

# Matriz de Confusion - test
matriz_conf=table(test$prediccion, test$Cluster_Aglom_FINAL)
prop.table(matriz_conf, 1) * 100

## Presición del modelo- test
print(100 * sum(diag(matriz_conf)) / sum(matriz_conf))
```

ANEXO II

```
###-------------- FACTOR DE AJUSTE ----------

y_nuevo_prueba_s3=y_construccion_S3[['Prediccion_Hueber_Seg3','INGRESO_C_NC']].copy()
dsvt=y_nuevo_prueba_s3['Prediccion_Hueber_Seg3'].std()


for factor in np.linspace(0.70,0.75,6):
    for f_desv in np.linspace(0.10,0.25,16):
        y_nuevo_prueba_s3['Ingreso_'+str(factor)+'_'+str(f_desv)] = np.where(y_nuevo_prueba_s3['Prediccion_Hueber_Seg3']>1000,y_nuevo_prueba_s3['Prediccion_Hueber_Seg3']*factor-f_desv*dsvt,y_nuevo_prueba_s3['Prediccion_Hueber_Seg3'])
        y_nuevo_prueba_s3['Ingreso_'+str(factor)+'_'+str(f_desv)+'_var']=(y_nuevo_prueba_s3['Ingreso_'+str(factor)+'_'+str(f_desv)]-y_nuevo_prueba_s3['INGRESO_C_NC'])/y_nuevo_prueba_s3['INGRESO_C_NC']
        print('Sobreestimacion_' + str(factor)+'_'+str(f_desv),len(y_nuevo_prueba_s3.loc[ y_nuevo_prueba_s3['Ingreso_'+str(factor)+'_'+str(f_desv)+'_var']>0.25,])/len(y_nuevo_prueba_s3))
        print('Subestimacion_' + str(factor)+'_'+str(f_desv),len(y_nuevo_prueba_s3.loc[ y_nuevo_prueba_s3['Ingreso_'+str(factor)+'_'+str(f_desv)+'_var']<-0.25,])/len(y_nuevo_prueba_s3))
       ## Graficas
        sns.distplot(y_nuevo_prueba_s3['INGRESO_C_NC'],color='g')
        sns.distplot(y_nuevo_prueba_s3['Prediccion_Hueber_Seg3'],color='y')
        sns.distplot(y_nuevo_prueba_s3['Ingreso_'+str(factor)+'_'+str(f_desv)],color='b')
        plt.show()
```

ANEXO III

El cálculo de las variables para el modelo de ingresos es de la siguiente manera:

- ✓ UTIL_MAXIMO_6UM: La utilización de línea de tarjeta de crédito en los últimos 6 meses se calcula:

$$\text{UTIL\_MAXIMO\_6UM}=\text{Máx}\left[\left(\frac{\text{SaldoNormal}}{\text{LineaNormal}}\right)*100\right]$$

- ✓ N_MESES_LINEA_TC: Meses con línea de tarjeta de crédito en el sistema financiero 12 meses es calculada a partir del monto de línea de la tarjeta de crédito en el sistema financiero.

  Primero se calcula un flag que indique como **1** si el cliente tiene un monto de línea mayor a cero.

$$\text{FLAG\_LINEA\_TC\_i}=\begin{cases} 1, & \text{CON\_TCRC\_LINEA\_TC}>0 \\ 0, & \text{caso contrario} \end{cases}$$

  Finalmente, la variable se construye como la sumatoria de los *flag's*.

$$\text{N\_MESES\_LINEA\_TC}=\sum_{i=1}^{12}\text{FLAG\_LINEA\_TC\_i}$$

- ✓ PAGO_TOTAL_6UM: El pago total en los 6 últimos meses, es la suma de los pagos realizados en cada mes por el cliente.

$$\text{PAGO\_TOTAL\_6UM} = \left(\text{Pago}_1+\text{Pago}_2+\ldots+\text{Pago}_{12}\right)$$

- ✓ MAX_NUMTC_CSALDO_12UM: El máximo número de tarjeta de crédito con saldo en 12 meses, es calculado a partir del número de tarjetas de crédito con saldo.

  En el caso de que el cliente tenga 2 códigos SBS entonces por ello se toma el máximo por cliente en un mismo periodo. Luego se tiene el número de tarjetas de crédito con saldo por periodo, entonces se procede a calcular el máximo en 12 meses.

$$\text{MAX\_NUMTC\_CSALDO\_12UM} =$$

$$\text{Máx(NUM\_TC\_SALDO\_1,NUM\_TC\_SALDO\_2, …, NUM\_TC\_SALDO\_12)}$$

- ✓ N_MESES_DISEF_TC: Meses con saldo disposición de efectivo en los 12 meses

Primero se calcula un *flag* que indique como **1** si el cliente tiene saldo en disposición de efectivo mayor a cero.

$$\text{FLAG\_CRED\_DISEF\_TC\_i}=\begin{cases} 1, & \text{CON\_TCRC\_DISP\_EFEC}>0 \\ 0, & \text{caso contrario} \end{cases}$$

Finalmente, la variable se construye como la sumatoria de los *flag's*.

$$\text{N\_MESES\_DISEF\_TC}=\sum_{i=1}^{12}\text{FLAG\_CRED\_DISEF\_TC\_i}$$

- ✓ PAGO_MEDIANA_6UM: Pago mediana en los últimos 6 meses

$$\text{PAGO\_TOTAL\_6UM} = \text{Mediana}(\text{Pago}_1, \text{Pago}_2, \ldots, \text{Pago}_{12})$$

- ✓ UTIL_PROM_6UM: Utilización de línea TC promedio en los 6 últimos meses

$$\text{UTIL\_PROM\_6UM} = \frac{\sum_{i=1}^{12} \left[\left(\frac{\text{SaldoNormal}_i}{\text{LineaNormal}_i} \right) * 100 \right]}{n}$$

donde n:número de meses donde tuvo utilización.

- ✓ Monto_tot_retail: Es el monto total de las transacciones realizadas en los 12 meses y además en los retail del grupo, tales como Saga, Tottus, Maestro y Sodimac.

$$\text{Monto\_tot\_retail} = (\text{Monto}_{\text{total}_{\text{Saga}}} + \text{Monto}_{\text{total}_{\text{Tottus}}} + \text{Monto}_{\text{total}_{\text{Sodimac}}} + \text{Monto}_{\text{total}_{\text{Maestro}}})$$

ANEXO IV

Cuadros resumen de lo modelos de Regresión Huber

Resumen Base Train	Segm. 1 (Ingresos Bajos)	Segm. 2 (Ingresos Medio-Alto)
Modelos Ingresos	Regresión Robusta (HUBER)	Regresión Robusta (HUBER)
n° Variables	7	6
R^2	0.11	0.13
Supuestos	Cumple todo	Cumple todo
Error Relativo	0.26	0.34
Sobreestimación 1	27.10%	36.20%
Factor Ajuste	Ingreso Estimado*0.90-0.15*Desv.Std.	Ingreso Estimado*0.90-0.15*Desv.Std.
Nuevo Error Relativo	0.23	0.29
Sobreestimación con Factor	11.50%	21.60%
Subestimación con Factor	28.50%	31.10%
Correcto con Factor	60.00%	47.20%

Resumen base Test	MODELO 1	
	Segm. 1 (Ingresos Bajos)	Segm. 2 (Ingresos Medio-Alto)
Modelos Ingresos	Regresión Robusta (HUBER)	Regresión Robusta (HUBER)
n° Variables	7	6
R^2	0.11	0.13
Supuestos	Cumple todo	Cumple todo
Error Relativo	0.25	0.34
Sobreestimación 1	26.80%	34.50%
Factor Ajuste	Ingreso Estimado*0.90-0.15*Desv.Std.	Ingreso Estimado*0.90-0.15*Desv.Std.
Nuevo Error Relativo	0.22	0.30
Sobreestimación con Factor	11.40%	20.30%
Subestimación con Factor	27.20%	32.20%
Correcto con Factor	61.40%	47.40%

REFERENCIAS

Hasih Pratiw, Y. S., & Twenty Liana4, S. (2014). M ESTIMATION, S ESTIMATION, AND MM ESTIMATION. *International Journal of Pure and Applied Mathematics, 91*(3), 349-360. doi:http://dx.doi.org/10.12732/ijpam.v91i3.7

Gallegos Gómez, J. L. (2008-2009). Apuntes de Econometría. LADE y LE.

Hothorn, T., Hornik , K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674.

ZAMAR, R. (1994). Estimación robusta. *ESTADISTICA ESPAÑOLA, 36*(137), 327-387.