

Modelo Ingreso No Cliente

Documento funcional de desarrollo

**MEMORIA METODOLÓGICA**

**Agosto 2019**

**DESCRIPCIÓN DEL DOCUMENTO**

|  |  |
| --- | --- |
| **Título** | **Modelo Ingreso No Cliente** |
| Descripción del documento | Documento funcional del desarrollo del modelo Ingresos No clientes con información del sistema financiero, transacciones y sociodemográficas. Detalla la metodología de construcción, las definiciones metodológicas y los resultados obtenidos. |
| Responsable de la documentación | **Corporativo Falabella Inversiones Financieras** |
| Fecha de Aprobación | **XXXXXX** |

**ACTUALIZACIONES Y MODIFICACIONES**

|  |  |  |  |
| --- | --- | --- | --- |
| Modificación efectuada | Realizado por | Fecha de realización | Fecha de aprobación |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Contenido**

[**1.** **Introducción** 5](#_Toc33536709)

[**2.** **Fuentes de Información** 5](#_Toc33536710)

[**3.** **Población Objetivo** 5](#_Toc33536711)

[**4.** **Periodo de Estudio** 6](#_Toc33536712)

[**5.** **Análisis Inicial de Datos** 6](#_Toc33536713)

[5.1. Análisis de Ingresos 7](#_Toc33536714)

[5.2. Análisis del Sistema Financiero 8](#_Toc33536715)

[5.3. Análisis de Transacciones 10](#_Toc33536716)

[5.4. Análisis Sociodemográfico 10](#_Toc33536717)

[**6.** **Tratamiento de Variables** 11](#_Toc33536718)

[6.1. Construcción de Variables 11](#_Toc33536719)

[6.1.1 Construcción de Variables RCC 11](#_Toc33536720)

[6.1.2 Construcción de Variable Transacciones 12](#_Toc33536721)

[6.1.3 Construcción de Variable Ubigeo 13](#_Toc33536722)

[6.2. Categorización de Variables y Valores Missing 13](#_Toc33536723)

[6.3 Análisis Multivariado y selección de variables 16](#_Toc33536724)

[**7.** **Construcción del Modelo Ingresos no Cliente** 17](#_Toc33536725)

[**8.** **Descripción del Modelo** 20](#_Toc33536726)

[**9.** **Desempeño de Modelos** 22](#_Toc33536727)

[ANEXOS 24](#_Toc33536728)

[I. Anexo (Creación de Variables) 24](#_Toc33536729)

[II. Anexo (Segmentación de Variables) 28](#_Toc33536730)

[III. Desempeño Segmentos 31](#_Toc33536731)

[1. Segmento Bajo (Regresión Cuantil) 31](#_Toc33536732)

[2. Segmento Medio (Regresión Cuantil) 31](#_Toc33536733)

[3. Segmento Alto (Regresión Cuantil) 31](#_Toc33536734)

[IV. Anexo Construcción de variables seleccionadas 32](#_Toc33536735)

[a. Saldo máximo en crédito de consumo de los últimos 12 meses (CON\_SALDO\_max) 32](#_Toc33536736)

[b. Reducción promedio mensual del saldo total en el sistema financiero. ( TOTAL\_SALDO\_cp\_mean2\_pos) 32](#_Toc33536737)

[c. Porcentaje de Utilización Máxima de Línea de crédito en los últimos 12 meses (CON\_TRX\_POR\_UTIL\_max) 33](#_Toc33536738)

[d. Mediana de los últimos 12 meses de la línea en tarjeta de crédito de la media por tarjeta en el sistema financiero (cupo\_tc\_mean\_median). 34](#_Toc33536739)

[e. Capacidad de compra dentro del retail (Frecuencia - Monto) 34](#_Toc33536740)

[f. Edad segmentada (Edad\_cat) 37](#_Toc33536741)

[g. Posesión de Vehículo (Flag\_Veh) 37](#_Toc33536742)

[h. Zona Geográfica 37](#_Toc33536743)

[V. Referencias Packages 38](#_Toc33536744)

[VI. Referencias 38](#_Toc33536745)

# **Introducción**

El modelo de Ingresos para no clientes nace con el propósito de generar un estimador robusto para la renta de potenciales clientes que puedan incorporarse a Banco Falabella Perú. El objetivo es controlar el potencial endeudamiento de un nuevo cliente, poniendo énfasis en la sobrestimación de la renta, evitando así, futuros riesgos de sobreendeudamiento.

Se presentan en este documento los aspectos metodológicos y los resultados obtenidos con respecto al proceso de construcción del modelo. En este contexto, es necesario abarcar los siguientes puntos: Análisis inicial de datos, creación y segmentación de variables, definición de metodologías y finalmente, el desempeño del modelo.

# **Fuentes de Información**

La información utilizada en la construcción del modelo de ingresos consta esencialmente de cuatro fuentes distintas:

1. Información del Reporte Crediticio Consolidado de Perú (RCC) para los años 2017 y 2018.
2. Información referente a transacciones realizadas por los clientes en los comercios de Saga Falabella, Tottus, Maestro y Sodimac para el periodo de enero a diciembre 2018.
3. Información sociodemográfica de los clientes consolidada a diciembre 2018.
4. Información de rentas es obtenida en referencia a las cuentas sueldo de Banco Falabella consolidadas a diciembre 2018.

Por otro lado, la extracción y validación de la información a utilizar en el modelo de Ingresos para No clientes ha sido realizada por el equipo de Data Quality de Banco Falabella Perú.

Las variables contenidas en cada una de las fuentes pueden observarse en el siguiente documento adjunto.



# **Población Objetivo**

La población objetivo del modelo de Ingresos para no clientes son aquellas personas naturales con al menos 1 Reporte Crediticio durante los 12 últimos meses.

Por otro lado, pese a que, las transacciones son una fuente de información importante en la construcción de este modelo, esta información es trabajada de forma tal que la participación de una persona dentro del retail no sea un motivo de exclusión. En el caso de nuevas admisiones, el ingreso estimado por medio de este modelo será vigente hasta que todas las condiciones necesarias para el modelo de ingreso cliente se cumplan.

# **Periodo de Estudio**

Para la definición de los periodos de estudios, el modelo se ajustó a la disponibilidad de la información.

1. Los ingresos corresponden a información de cuentas sueldo consolidadas a diciembre de 2018. Por lo tanto, toda la información utilizada para medir el comportamiento del cliente mantiene como referencia este periodo.
2. Con el objetivo de utilizar la información más reciente a la fecha con una temporalidad de 12 meses y considerando un retraso de 2 meses en la recepción de la información, las bases de datos correspondientes al RCC son las existentes entre noviembre 2017 y octubre 2018.
3. Por otro lado, la información interna de transacciones de retail son desde enero a diciembre 2018, en referencia a los 12 meses más recientes de información.
4. Finalmente, la información sociodemográfica corresponde a un consolidado disponible a diciembre de 2018.

Ilustración 1: Periodos de estudio

# **Análisis Inicial de Datos**

Con el objetivo de conocer el universo de variables disponibles, comprendiendo el comportamiento de distribución y la presencia de casos atípicos se realizó un análisis inicial de datos de todas las fuentes de información disponibles:

1. Ingresos cuenta sueldo,
2. Información del Sistema Financiero (RCC),
3. Transacciones dentro del grupo e
4. Información sociodemográfica.

## Análisis de Ingresos

Corresponde a la variable respuesta del modelo. Se tiene información de 49.777 personas, quienes tienen un ingreso promedio de 2.098 soles con una desviación estándar de 1.902 soles. Se observa una distribución con asimetría positiva, donde el ingreso máximo observado es de 53.345 soles.

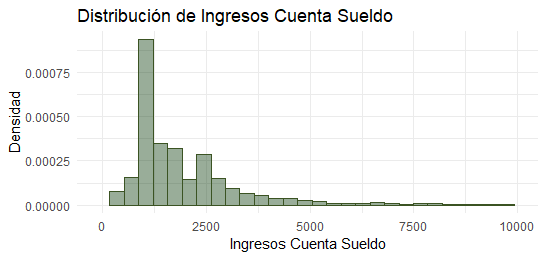
Tabla 1 Distribución de Ingresos

Ilustración 2:Distribución de ingresos

|  |  |  |
| --- | --- | --- |
| **Indicadores** | | **Valores** |
| N | | 49.777 |
| Desv.Est. | | 1.902 |
| Media | | 2.098 |
| Percentil | 0% | 0,01 |
| 10% | 905 |
| 20% | 981 |
| 30% | 990 |
| 40% | 1.208 |
| 50% | 1.486 |
| 60% | 1.784 |
| 70% | 2.433 |
| 80% | 2.807 |
| 90% | 3.927 |
| 100% | 53.345 |

Con el fin de obtener estimaciones robustas sin verse afectadas por valores extremos, atípicos y/o no confiables, se consideran los siguientes filtros referentes a las fuentes de ingreso al momento de la construcción:

Tabla 2 Filtros de Ingresos para la construcción

|  |  |  |  |
| --- | --- | --- | --- |
| **Filtro** | **Base Inicial** | **Base Final** | **Registros Eliminados** |
| Se eliminan los registros de cuenta sueldo para trabajadores independientes debido a la baja confiabilidad de sus informaciones. | 49.777 | 49.041 | 736 |
| Se elimina el 5% de ingresos extremos (2,5% de los ingresos más bajos y 2,5% de los ingresos más altos), con el objetivo de evitar valores atípicos. | 49.041 | 46.550 | 2.491 |
| Debido a la escasa fuente de información sociodemográfica, se elimina todo número de documento que no haya tenido reporte en el RCC en los últimos 12 meses. | 46.550 | 37.870 | 8.680 |

*Todos estos filtros son a nivel metodológico para evitar inconvenientes con el ajuste de los modelos, por lo mismo, éstos son aplicados posterior a finalizar el proceso de construcción de variables.*

## Análisis del Sistema Financiero

Como primera consideración al momento de analizar la información del sistema financiero, resulta necesario eliminar los casos duplicados de códigos SBS y/o número de documentos nulo correspondiente a personas jurídicas.

Tabla 3 Filtros Sistema Financiero

|  |  |  |  |
| --- | --- | --- | --- |
| **Filtro** | **Base Inicial** | **Base Final** | **Registros Eliminados** |
| Número de documento nulos (personas jurídica) y registros con distinto código SBS e igual número y tipo de documento. | 2017: 117.276.142  2018: 122.686.789 | 2017: 114.247.167  2018: 119.526.686 | 2017: 3.028.975  2018: 3.160.103 |

Para esta fuente de datos, se utilizan únicamente los 12 meses de información observables más recientes previas a la medición de la renta (los meses comprendidos entre noviembre 2017 a octubre 2018, ambos incluidos). Se cuenta con un total de 118 variables relacionadas a diferentes aspectos como moras, saldos, créditos, entre otros. El total de variables comprendidas en el sistema financiero puede revisarse en el siguiente documento adjunto.



A modo de resumen, se muestran estadísticos descriptivos de las variables que por criterio de negocio han sido seleccionadas como principales factores influyentes en un estimador de renta.

Tabla 4 Análisis descriptivo RCC

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **N** | | **Media** | **Desviación estándar** | **Mínimo** | **Percentiles** | | | | | **Máximo** |
| **Válido** | **Perdidos** | **5** | **25** | **50** | **75** | **95** |
| **CON\_CCNR\_SALDO** | 453.846 | 143.478 | 2.654,24 | 28.574,86 | - | - | - | - | 500,00 | 13.716,95 | 5.344.834,03 |
| **CON\_CCRV\_LINEA\_NO\_UTIL** | 453.846 | 143.478 | 21,11 | 227,56 | - | - | - | - | - | - | 17.476,95 |
| **CON\_CCRV\_SALDO** | 453.846 | 143.478 | 13,86 | 445,30 | - | - | - | - | - | - | 79.991,85 |
| **CON\_CCTE\_SALDO** | 453.846 | 143.478 | 4,21 | 288,36 | - | - | - | - | - | - | 35.475,54 |
| **CON\_CVEH\_SALDO** | 453.846 | 143.478 | 789,53 | 5.356,04 | - | - | - | - | - | - | 146.800,00 |
| **CON\_SALDO** | 453.846 | 143.478 | 8.447,19 | 33.066,22 | - | - | 0,92 | 1.548,33 | 8.160,70 | 38.811,60 | 5.431.409,04 |
| **CON\_TCRC\_COMPRAS** | 453.846 | 143.478 | 1.922,79 | 5.122,70 | - | - | - | 471,83 | 1.886,60 | 8.077,74 | 268.731,63 |
| **CON\_TCRC\_COMPRAS\_VIG** | 453.846 | 143.478 | 1.816,55 | 4.788,71 | - | - | - | 427,47 | 1.796,11 | 7.715,57 | 268.731,63 |
| **CON\_TCRC\_DISP\_EFEC** | 453.846 | 143.478 | 670,53 | 3.117,47 | - | - | - | - | - | 3.525,95 | 116.057,78 |
| **CON\_TCRC\_OTROS** | 453.846 | 143.478 | 2.182,91 | 7.945,52 | - | - | - | - | - | 12.900,18 | 241.456,19 |
| **CON\_TCRC\_SALDO** | 453.846 | 143.478 | 4.776,22 | 12.008,36 | - | - | - | 737,47 | 4.223,89 | 22.593,81 | 356.132,09 |
| **CONT\_SALDO** | 453.846 | 143.478 | 0,33 | 105,98 | - | - | - | - | - | - | 49.000,00 |
| **cupo\_medio\_tc** | 343.083 | 254.241 | 6.977,54 | 9.166,13 | - | 500,00 | 1.249,02 | 3.785,36 | 8.950,01 | 24.573,60 | 183.796,26 |
| **cupo\_tc** | 453.846 | 143.478 | 13.599,55 | 28.372,69 | - | - | 1,00 | 3.000,00 | 14.500,02 | 60.919,25 | 551.388,77 |
| **HIP\_SALDO** | 453.846 | 143.478 | 4.685,70 | 40.019,42 | - | - | - | - | - | - | 4.345.564,14 |
| **INTS\_SALDO** | 453.846 | 143.478 | 131,95 | 479,84 | - | - | - | 15,22 | 117,90 | 596,31 | 56.707,60 |
| **MAX\_CLASIFICACION** | 431.122 | 166.202 | 1,23 | 1,78 | - | - | - | - | 4,00 | 4,00 | 4,00 |
| **MIC\_SALDO** | 453.846 | 143.478 | 318,74 | 4.748,11 | - | - | - | - | - | - | 392.550,78 |
| **NEG\_SALDO** | 453.846 | 143.478 | 45,89 | 4.109,93 | - | - | - | - | - | - | 509.285,29 |
| **NUM\_ENTIDADES** | 453.846 | 143.478 | 2,29 | 1,37 | 1,00 | 1,00 | 1,00 | 2,00 | 3,00 | 5,00 | 14,00 |
| **NUM\_ENTIDADES\_SALDO** | 453.846 | 143.478 | 1,88 | 1,18 | - | 1,00 | 1,00 | 2,00 | 2,00 | 4,00 | 13,00 |
| **NUM\_HIPO** | 453.846 | 143.478 | 0,03 | 0,19 | - | - | - | - | - | - | 2,00 |
| **NUM\_PREST** | 453.846 | 143.478 | 0,40 | 0,66 | - | - | - | - | 1,00 | 2,00 | 10,00 |
| **NUM\_TC** | 453.846 | 143.478 | 1,59 | 1,46 | - | - | 1,00 | 1,00 | 2,00 | 4,00 | 10,00 |
| **NUM\_TC\_SALDO** | 453.846 | 143.478 | 1,15 | 1,15 | - | - | - | 1,00 | 2,00 | 3,00 | 8,00 |
| **NUM\_VEH** | 453.846 | 143.478 | 0,03 | 0,18 | - | - | - | - | - | - | 3,00 |
| **TOTAL\_SALDO** | 453.846 | 143.478 | 15.664,94 | 65.007,87 | - | 2,17 | 923,35 | 3.518,29 | 12.708,55 | 63.974,01 | 9.791.752,49 |

Los resultados expuestos en la tabla anterior hacen referencia al comportamiento en el sistema financiero de las 49.777 cuentas sueldo para los 12 meses mencionados en los puntos anteriores, dejando un total de 597.324 registros por variable. En la tabla queda plasmado que 143.478 observaciones no tienen visualizaciones en el sistema financiero.

El análisis descriptivo de todas las variables del sistema financiero para el periodo considerado puede observarse en el siguiente documento adjunto.



## Análisis de Transacciones

Como tercera fuente de información se encuentran las transacciones en los *retail’s* del grupo Falabella, las cuales por protocolo del sistema son recopiladas al momento en que un cliente proporciona su DNI al pagar con tarjetas de crédito dentro de las tiendas de Saga Falabella, Tottus, Sodimac y Maestro. Pese a que, estas fuentes contienen información detallada de la compra: cantidades, tipos de productos, descripción y categoría de productos… las transacciones serán analizadas a nivel DNI, considerando recurrencias e importe total en los *retail’s*. Los resultados se muestran en la siguiente tabla:

Tabla 5 Análisis Descriptivo Transacciones

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **N** | | **Media** | **Desviación Estándar** | **Mínimo** | **Percentil** | | | | | **Máximo** |
| **Validos** | **Perdidos** | **5%** | **25%** | **50%** | **75%** | **95%** |
| **Importe Maestro** | 894.146 | 1.915.443 | 683,33 | 4.241,70 | 0,27 | 11,60 | 54,29 | 185,70 | 563,96 | 2.498,64 | 1.645.163,00 |
| **Número Compras Maestro** | 894.146 | 1.915.443 | 2,95 | 4,47 | 1,00 | 1,00 | 1,00 | 2,00 | 3,00 | 9,00 | 273,00 |
| **Importe Sodimac** | 1.405.764 | 1.403.825 | 706,72 | 5.336,52 | 0,30 | 14,90 | 68,60 | 233,60 | 677,94 | 2.557,00 | 4.022.504,00 |
| **Número Compras Sodimac** | 1.405.764 | 1.403.825 | 3,22 | 4,28 | 1,00 | 1,00 | 1,00 | 2,00 | 4,00 | 10,00 | 263,00 |
| **Importe Tottus** | 1.721.773 | 1.087.816 | 870,00 | 2.949,45 | 0,30 | 28,70 | 131,37 | 384,45 | 1.010,68 | 3.132,78 | 1.030.808,00 |
| **Número de Compras Tottus** | 1.721.773 | 1.087.816 | 6,62 | 9,69 | 1,00 | 1,00 | 1,00 | 3,00 | 8,00 | 23,00 | 309,00 |
| **Importe**  **Saga** | 1.262.232 | 1.547.357 | 1.543,93 | 4.748,03 | 0,01 | 59,00 | 249,00 | 720,00 | 1.869,70 | 5.479,70 | 2.799.306,00 |
| **Número de Compras**  **Saga** | 1.262.232 | 1.547.357 | 5,41 | 6,41 | 1,00 | 1,00 | 1,00 | 3,00 | 7,00 | 18,00 | 363,00 |

Se observa que del total de 2.808.589 clientes que entregaron el número de DNI al momento de una compra en los comercios considerados. La mayor participación es encontrada en Tottus, teniendo un importe medio anual de 870 soles por persona. Por otro lado, Saga Falabella presenta el importe medio anual por cliente más alto entre los cuatro *retail’s* llegando a 1.547 soles.

## Análisis Sociodemográfico

Finalmente, la información sociodemográfica contiene un total de 25.457.476 de personas observadas, en donde se almacena información referente a género, edad, estado civil, ubigeo, grado de institución, nivel socioeconómico e información respecto a vehículos inscritos. Al ser una fuente de información sin mucha variabilidad mensual, este reporte es considerado como un consolidado a diciembre 2018.

# **Tratamiento de Variables**

El siguiente apartado tiene por objetivo explicar los procesos a los cuales fueron sometidos las variables iniciales exhibidas en los puntos anteriores. Estos procesos se realizaron con el objetivo de neutralizar los posibles problemas que pueden generar la presencia de *missings*, *outliers* o acumulación de variables en un punto. A continuación, se detalla el proceso de categorización de las variables predictoras (en caso de ser necesario) y los análisis bivariados correspondientes.

## Construcción de Variables

La estrategia a utilizar en la construcción de variables se encuentra relacionada en gran parte a la fuente de información y la estructura de las variables disponibles, incorporando un horizonte de 12 meses de observación, en el caso de ser posible. Cabe destacar que, debido a la estructura y disponibilidad de la información sociodemográfica, en esta fuente de información únicamente se genera la transformación de fecha de nacimiento a edades (en años), tomando como referencia el 31 de diciembre de 2018.

### Construcción de Variables RCC

Como primer paso en la construcción de variables del sistema financiero, de las 118 variables disponibles detalladas en el apartado **5.2**, se configuró un conjunto de 30 variables. Éstas fueron seleccionadas dado que resultan ser las más representativas del total.

Tabla 6 Construcción de Variables RCC

|  |  |  |
| --- | --- | --- |
| **N** | **Variables** | **Descripción** |
| 1 | RCC | Numero de visualizaciones en el sistema financiero |
| 2 | LINEA\_TC\_MEAN | Monto de línea promedio por tarjeta de crédito |
| 3 | CON\_TCRC\_LINEA\_TC | Monto de línea de la tarjeta de crédito |
| 4 | CON\_TRX\_POR\_UTIL | Porcentaje Utilizado de la línea de crédito |
| 5 | CON\_TRX\_NOUTIL | Porcentaje No Utilizado de la línea de crédito |
| 6 | MAX\_CLASIFICACION | Máxima clasificación en el periodo |
| 7 | NUM\_ENTIDADES | Número de entidades reportantes |
| 8 | NUM\_ENTIDADES\_SALDO | Número de entidades con saldo |
| 9 | NUM\_TC | Número de tarjetas de crédito (entidades) |
| 10 | NUM\_TC\_SALDO | Número de tarjetas de crédito con saldo |
| 11 | NUM\_PREST | Número de préstamos |
| 12 | NUM\_VEH | Número de créditos vehiculares |
| 13 | NUM\_HIPO | Número de créditos hipotecarios |
| 14 | TOTAL\_SALDO | Monto total de saldo en el SF en el mes |
| 15 | NEG\_SALDO | Monto de saldo en créditos comerciales |
| 16 | MIC\_SALDO | Monto de saldo en créditos de micro y pequeña empresa |
| 17 | CON\_SALDO | Monto de saldo en créditos de consumo |
| 18 | CON\_TCRC\_SALDO | Monto de saldo en créditos revolventes de consumo con tarjeta de crédito |
| 19 | CON\_TCRC\_COMPRAS | Monto de saldo de compras en créditos de consumo con tarjeta de crédito |
| 20 | CON\_TCRC\_DISP\_EFEC | Monto de saldo de disposición de efectivo en créditos de consumo con tarjeta de crédito |
| 21 | CON\_TCRC\_OTROS | Monto de saldo en otros de tarjeta de crédito |
| 22 | CON\_TCRC\_COMPRAS\_VIG | Monto del saldo en compras vigente |
| 23 | CON\_CCRV\_SALDO | Monto del saldo en créditos de consumo revolventes |
| 24 | CON\_CCRV\_LINEA\_NO\_UTIL | Línea no utilizada de consumos revolventes |
| 25 | CON\_CCNR\_SALDO | Monto del saldo en créditos de consumo no revolventes |
| 26 | CON\_CVEH\_SALDO | Monto del saldo en créditos de consumo vehiculares |
| 27 | CON\_CCTE\_SALDO | Monto del saldo en créditos de consumo cuenta corriente |
| 28 | HIP\_SALDO | Monto del saldo en créditos hipotecarios |
| 29 | INTS\_SALDO | Monto de intereses |
| 30 | CONT\_SALDO | Saldo contingente |

Para cada una de las variables expuesta en la **Tabla 6** se procedió a resumir la información observada de los 12 meses por estadísticos de posición, escala y de tendencia como son: el máximo, mínimo, media, mediana, variación mensual, entre otros. En resumen, para cada variable se midió un total de 18 versiones, originando un total de 540 variables referente al RCC.



Para analizar en detalle este proceso se recomienda dar revisión al [**Anexo I**](#_Anexo_(Creación_de)**.**

### Construcción de Variable Transacciones

Con el objetivo de medir la capacidad adquisitiva de cada persona, las transacciones disponibles en retail fueron analizadas en base a 2 directrices: Por un lado, se analiza el número de compras que ha realizado el cliente en los últimos 12 meses disponibles, y, por otro lado, se mide el importe total en soles que el cliente ha desembolsado en el retail durante los últimos 12 meses observados. Estas mediciones se hacen a modo individual para de Tottus, Saga, Sodimac y Maestro, y también a modo consolidado.

### Construcción de Variable Ubigeo

Con el objetivo de generar variables consistentes en base a la información disponible, los ubigeos son reagrupados según proximidad geográfica (conocimiento a priori del negocio en base a las rentas estimadas por zonas geográficas), en donde la variable final consta únicamente de 3 zonas. Los detalles de los ubigeos que contiene cada zona geográfica pueden ser revisados en el siguiente documento adjunto.



## Categorización de Variables y Valores Missing

Debido a la alta acumulación de **0** en algunas variables asociadas al Sistema Financiero (tal como se detalla en el apartado **5.2**) y con el objetivo de incorporar relaciones no lineales con el ingreso, toda variable continua es categorizada. Para esto, las variables resultantes del sistema financiero, en conjunto con la edad y el grado de instrucción dentro de la información sociodemográfica, son segmentadas en un proceso iterativo, en donde la información observada es fraccionada por medios de árbol de clasificación con metodología CART, incorporando las categorías de **missing** y **0** de manera independiente con el objeto de comprender su naturaleza y reagruparlas correctamente. Los resultados de esta etapa pueden observarse en el siguiente documento:



Todos los procesos de categorización fueron ejecutados por medio del programa *RStudio* y el desarrollo correspondiente a nivel código y metodológico puede ser revisado a detalle en el [**Anexo II**](#_Anexo_(Segmentación_de). Posterior a la categorización encontrada por el algoritmo, las variables son sometidas a un último filtro y manipulación manual, en donde se agrupan las categorías con sentidos similares y baja concentración de población, respetando también que cada variable mantenga un sentido lógico y alineado con el negocio.

Ejemplos de este proceso puede observarse en la siguiente tabla.

Tabla 7 Segmentación de Variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Categoría** | **N** | **N %** | **N % sin perdidos** | **Ingreso Promedio** | **Beta Regresión Lineal Simple** |
| TOTAL\_SALDO\_cp\_mean2\_pos | <= 577.77 | 18.495 | 37% | 55% | 1.771 | 1.771 |
| TOTAL\_SALDO\_cp\_mean2\_pos | 577.77<·<=1328.0457 | 8.173 | 16% | 25% | 2.467 | 696 |
| TOTAL\_SALDO\_cp\_mean2\_pos | 1328.0457<·<=3630.73 | 4.959 | 10% | 15% | 3.386 | 1.615 |
| TOTAL\_SALDO\_cp\_mean2\_pos | > 3630.73 | 1.701 | 3% | 5% | 4.897 | 3.126 |
| TOTAL\_SALDO\_cp\_mean2\_pos | Missing | 9.213 | 19% | 28% | 1.505 | -266 |
| TOTAL\_SALDO\_cp\_mean2\_pos | Missing\_var | 7.236 | 15% | 22% | 1.735 | -36 |
| CON\_SALDO\_max | 0 | 7.880 | 16% | 19% | 1.762 | 1.762 |
| CON\_SALDO\_max | <=5565.55 | 16.937 | 34% | 42% | 1.834 | 72 |
| CON\_SALDO\_max | 5565.55<·<=19809.84 | 9.102 | 18% | 22% | 2.442 | 680 |
| CON\_SALDO\_max | 19809.84<·<=43697.65 | 4.205 | 8% | 10% | 3.029 | 1.267 |
| CON\_SALDO\_max | >43697.65 | 2.440 | 5% | 6% | 4.384 | 2.622 |
| CON\_SALDO\_max | Missing | 9.213 | 19% | 23% | 1.505 | -257 |
| Edad | (-Inf,25] | 11.140 | 22% | 22% | 1.333 | 1.333 |
| Edad | (25,27] | 4.620 | 9% | 9% | 1.754 | 421 |
| Edad | (27,32] | 10.668 | 21% | 21% | 2.143 | 810 |
| Edad | (32,Inf] | 23.349 | 47% | 47% | 2.512 | 1.179 |

*En donde las categorías* ***Missing*** *corresponde a las observaciones que no poseen información en su origen y missing\_var son aquellas que teniendo información no cumplen con las condiciones para ser calculadas, como cocientes o variaciones intermensuales cuando solo se tiene información de 1 mes.*

Por otro lado, y debido a los altos niveles de correspondencia que existe entre el importe total y en número de compras que realiza un cliente dentro el retail, estas variables tuvieron que ser integradas con el objetivo de evitar futuros problemas de dependencia entre variables explicativas.

En primer lugar, los importes y número de compras fueron considerados a nivel consolidado sin hacer distinción entre las tiendas en que el cliente ha realizado la compra. El proceso de segmentación sigue los siguientes pasos: tanto la variable de importe total como el número de compras realizadas es segmentada en 9 categorías por medio de deciles, en donde la primera categoría se ve concentrada por todas las personas que no realizan transacciones dentro del grupo:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Importe total realizado en los últimos 12 meses de observación | | | Número de compras totales realizadas en los últimos 12 meses de observación | | |
| Segmento | Número de Casos | Porcentaje en la población | Segmento | Número de Casos | Porcentaje en la población |
| (-1,0] | 13.510 | 27% | (-1,0] | 13510 | 27% |
| (0,28.7] | 1.425 | 9% | (0,1] | 4533 | 3% |
| (28.7,204] | 4.976 | 5% | (1,2] | 2717 | 10% |
| (204,603] | 4.978 | 10% | (2,5] | 4911 | 10% |
| (603,1.270] | 4.977 | 10% | (5,10] | 4948 | 10% |
| (1.270,2.290] | 4.978 | 10% | (10,17] | 4576 | 10% |
| (2.290,3.890] | 4.977 | 10% | (17,28] | 4627 | 10% |
| (3.890,7.020] | 4.978 | 10% | (28,50] | 5073 | 10% |
| (7.020,...] | 4.978 | 10% | (50,…] | 4882 | 10% |

Tabla 8 Importe y compras retail

Teniendo cada dimensión con 9 categorías, se procede a unificar los importes y los números de compras en una sola variable de frecuencia y monto, la cual hace referencia a la categoría mínima a la que pertenece el cliente, entre el importe realizado y el número de veces que ha comprado dentro del retail del grupo Falabella.

Tabla 9 Variable mixta Frecuencia - Monto

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Número de Compras | | | | | | | | | Variable Mixta Frecuencia Monto |
|  |  | (-1,0] | (0,1] | (1,2] | (2,5] | (5,10] | (10,17] | (17,28] | (28,50] | (50,…] |  |
| Monto Total Importado | (-1,0] | 27,14% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 1 |
| (0,28.7] | 0,00% | 2,55% | 0,27% | 0,04% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 2 |
| (28.7,204] | 0,00% | 4,61% | 2,67% | 2,29% | 0,38% | 0,05% | 0,00% | 0,00% | 0,00% | 3 |
| (204,603] | 0,00% | 1,42% | 1,74% | 3,86% | 2,22% | 0,64% | 0,10% | 0,02% | 0,00% | 4 |
| (603,1.270] | 0,00% | 0,40% | 0,57% | 2,36% | 3,49% | 2,09% | 0,82% | 0,24% | 0,03% | 5 |
| (1.270,2.290] | 0,00% | 0,10% | 0,14% | 0,90% | 2,36% | 2,98% | 2,26% | 1,13% | 0,14% | 6 |
| (2.290,3.890] | 0,00% | 0,03% | 0,06% | 0,32% | 1,03% | 2,14% | 3,13% | 2,52% | 0,78% | 7 |
| (3.890,7.020] | 0,00% | 0,00% | 0,01% | 0,09% | 0,38% | 1,06% | 2,29% | 3,71% | 2,46% | 8 |
| (7.020, ...] | 0,00% | 0,00% | 0,00% | 0,02% | 0,08% | 0,23% | 0,69% | 2,58% | 6,40% | 9 |
| Variable Mixta Frecuencia Monto | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |

Finalizada la creación y posterior segmentación de variables, se procede a medir el poder discriminante ajustando regresiones lineales simples, utilizando como variable respuesta la renta y como predictor la versión categorizada de cada variable. Posteriormente, se miden estadísticos como R2, R2 ajustado, AIC y BIC, seleccionando variables con R2 ajustado de un mínimo de 5% que mantengan buenos indicadores de AIC y BIC.

El set final de variables seleccionadas para el modelo puede ser revisado en el documento adjunto:



## Análisis Multivariado y selección de variables

En esta sección se describe el análisis multivariado enfocado en análisis de correlaciones por medio del estadístico V de Cramer, el cual es importante para prevenir futuros problemas de multicolinealidad en el modelo. Ésta es una medida simétrica para la intensidad de la relación entre dos o más variables de la escala nominal, cuando al menos una de las dos variables tiene por lo menos dos valores posibles. Su rango es [0, 1], donde el 1 significa que hay una correlación perfecta entre las variables, y 0 significa que no existe correlación. Según Lineamientos de Construcción Modelos de Crédito, se considera que la correlación V de Cramer entre 2 variables independientes es intensa si supera el 0.4, y es muy fuerte (inaceptable) si supera el 0.6.

En resumen, la selección de variables finales para el modelo considera lo siguiente:

1. Las variables seleccionadas mantienen una tendencia acorde con el negocio.
2. Con el fin de evitar problemas de multicolinealidad y cumpliendo con los lineamientos de Modelos, las variables seleccionadas deben tener como máximo una correlación de V de Cramer como máximo de 0.6.
3. Las variables deben poseer un alto nivel predictivo a nivel individual, con un R2 ajustado como mínimo de un 5% e índices de AIC y BIC lo más bajo posible.

En el archivo adjunto, se muestra el análisis de correlación aplicado en la selección de variables realizadas en el punto 6.2.



# **Construcción del Modelo Ingresos no Cliente**

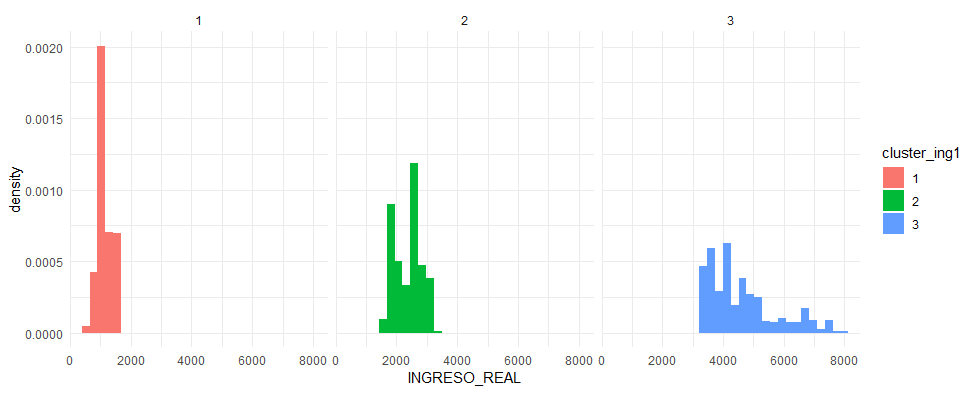
Para la construcción del Modelo de Ingresos ha sido necesario la utilización de 2 metodologías de estimación concatenadas en un único proceso. En primer lugar, Modelos Logísticos Multinomiales seguido de Regresión cuantil a la mediana.

Debido a la asimetría que presentan los ingresos dentro de la población, el primer modelo tiene por objetivo discernir los diferentes segmentos de renta con el objetivo de controlar los efectos de sobrestimación y subestimación. Para lograr esto, se definen los segmentos de renta por medio de cluster de k-medoides utilizando el algoritmo Clara (optimización del algoritmo PAM, *partitioning around medoids*), un método de agrupamiento que tiene por objetivo particionar el espacio en **k** grupos, cuya medida de distancia es minimizada inter grupo. A nivel estratégico se desarrolla el algoritmo para discriminar 3 segmentos de ingreso.

Tabla 9 Definición de Cluster

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Clúster** | **Número de casos** | **Proporción de la población** | **Ingreso Medio** | **Ingreso Mínimo** | **Ingreso Máximo** |
| 1.- Ingresos Bajos | 19.735 | 52,1% | 1.127 | 516 | 1.639 |
| 2.- Ingresos Medios | 13.095 | 34,6% | 2.356 | 1.639 | 3.240 |
| 3.- Ingresos Altos | 5.040 | 13,3% | 4.611 | 3.245 | 7.999 |

Ilustración 2 Distribución de los Ingresos según Cluster



Definimos los rangos Bajo, Medio y Alto para la renta, se procede a ajustar modelo Logístico Multinomial con el objetivo de estimar el segmento al cual pertenece cada persona.

Sea  la probabilidad de que un sujeto pertenezca al segmento y el conjunto de variable independientes disponibles para el modelo. Luego,



El modelo Logístico Multinomial tiene por objetivo la estimación de las probabilidad de pertenecer a cada segmento cumpliendo la restricción que .

Debido a esta última restricción, es que la ejecución del modelo multinomial define en su origen una categoría base, en particular para este caso la categoría base corresponderá a la central . Finalmente, se definen los modelos logísticos en base al logaritmo de la razón de probabilidades que tiene un segmento con respecto al base:

Donde y con son los parámetros de peso a estimar por el modelo.

Luego, la probabilidad de que un sujeto pertenezca al segmento 1, 2 o 3 queda determinado por:

Finalmente, la estimación del modelo logístico multinomial queda determinado por el segmento con mayor probabilidad.

Posterior a la definición de intervalo de renta al cual pertenece el cliente, se realiza una estimación puntual, condicionada al segmento estimado por la regresión multinomial. Como resultado de una serie de iteraciones en la construcción de modelo y teniéndose la asimetría de los datos, la metodología utilizada para la estimación puntual de la renta es mediante regresiones cuantil a la mediana (percentil 50).

La regresión Cuantil es introducida por Koenker & Bassett en 1978, los cuales proponen la estimación de los parámetros de peso por medio de la reducción del error en base a una función de error que diverge de mínimos cuadrados y centra la optimización de los parámetros en la reducción del error para un percentil objetivo, es decir:

En donde hace hace referencia a la matriz de diseño y b es el vector de parámetros de peso compuesto por . El algoritmo para la estimación de los parámetros de peso corresponde a Frisch-Newton desarrollado por Portnoy & Koenker en 1997.

De forma general, en las 2 etapas de construcción del modelo se ha realizado una selección de variables significativas, intentando cumplir el principio de parsimonia (es decir, que aporten la mayor capacidad discriminante con el menor número de variables). Debido al cambio de población que existe entre el modelo inicial de segmentación y el segundo modelo de estimación puntual las variables fueron sometidas a un reordenamiento, agrupando las categorías que no contenían un tamaño suficiente para tener una relevancia significativa.

Durante el proceso de selección de variables, se realizaron diferentes modelos tentativos, procurando alcanzar un modelo final que sea estadísticamente aceptable, cumpliendo los siguientes criterios:

* Lógica de la beta correctos: signo coherente con el sentido económico de la variable, mantiene la tendencia inicial de la misma.
* Coeficientes de regresión significativos: el p-valor del test de hipótesis que evalúa la significancia de los coeficientes, debe ser menor al 5% (según lineamientos corporativos). Lo que busca esta prueba es el rechazo de la hipótesis nula definida como H0: . En caso de ser superior a 0.05, se da por no rechazada y esto indica que la categoría de la variable es no significativa para la clasificación del modelo.
* Peso de las variables: El peso de una variable es determinado por la fracción de desviación estándar en el modelo que es inducida por cada variable. Con el objetivo de que todas las variables aporten equitativamente a la estimación de la renta, se verificó que no exista concentración en las variables que conforman el modelo, seleccionado únicamente las variables que tuvieran un peso mayor al 5%.

Finalmente, se realiza una validación cruzada, en donde la construcción y estimación de los parámetros es generada en una muestra del 70% para la construcción y la validación en base al 30% restante, para un total de 100 repeticiones. Utilizando la metodología *bootstrap* por medio del estadístico de media de todas las iteraciones, la obtención de un parámetro que sea consistente para todo subconjunto de la población estudiada sin estar condicionado a una única muestra de construcción.

A modo de resumen el proceso de estimación puede ser sintetizado en el siguiente cuadro.

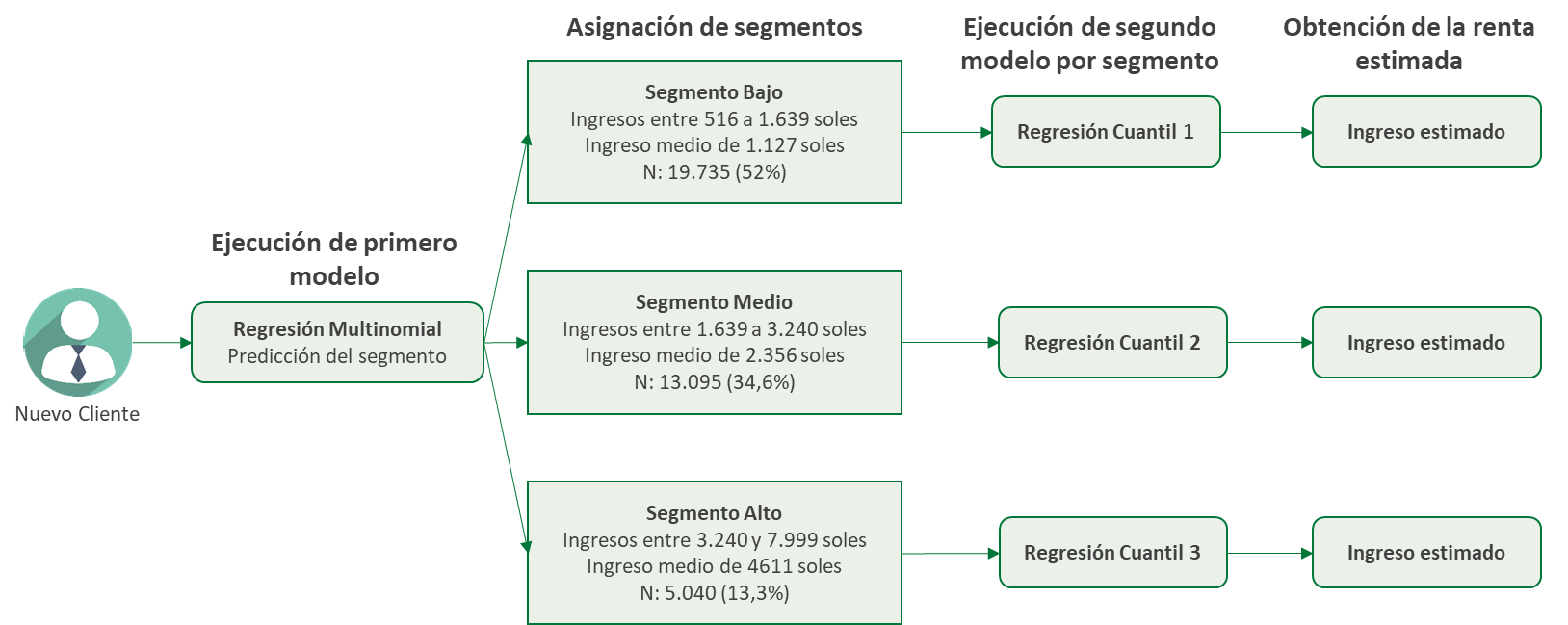
****

Ilustración 3 Flujo Metodológico de estimación

# **Descripción del Modelo**

Una vez definida la metodología de construcción y validación, el desarrollo y estimación de los parámetros fue realizado en el programa de *RStudio* en base a los paquetes de **nnet** para la regresión de logística multinomial y **quantreg** para ajustar las regresiones cuantiles a la mediana. En la búsqueda del mejor modelo y en cumplimiento de todas las restricciones de selección de variables comentadas en los puntos anteriores, algunas variables son recategorizadas y/o eliminadas. Los resultados finales para el modelo de segmentación se observan en lasiguientetabla.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Intervalo** | **Beta 1** | **Beta 3** | **Valor-p Beta 1** | **Valor-p Beta 3** | **Peso Beta 1** | **Peso Beta 3** |
| (Intercept) |  | 1,460 | -2,158 |  |  |  |  |
| CON\_SALDO\_max | (5.565,55 a 19.809,84] | -0,225 | -0,307 | <0,001 | <0,001 | 11% | 8% |
| (19.809,84 a 43.697,65] | -0,508 | -0,242 | <0,001 | <0,001 |
| Mayor a 43.697,65 | -0,650 | 0,169 | <0,001 | <0,001 |
| TOTAL\_SALDO\_cp\_mean2\_pos | (577.77; 1.328,04571428571] | -0,352 | 0,260 | <0,001 | <0,001 | 14% | 15% |
| (1.328,04571428571; 3.630,73] | -0,525 | 0,665 | <0,001 | <0,001 |
| Mayor a 3.630,73 | -0,904 | 1,038 | <0,001 | <0,001 |
| CON\_TRX\_POR\_UTIL\_max | (0,20; 0,365] | 0,039 | -0,043 | <0,001 | <0,001 | 8% | 14% |
| (0,365; 0,694] | 0,355 | -0,541 | <0,001 | <0,001 |
| Mayor a 0.694 | 0,283 | -0,500 | <0,001 | <0,001 |
| cupo\_tc\_mean\_median | (3.495,75; 8.413,92] | -0,396 | 0,257 | <0,001 | <0,001 | 18% | 16% |
| (8.413,92; 18.975,01] | -0,821 | 0,550 | <0,001 | <0,001 |
| Mayor a 18.975,01 | -0,859 | 1,185 | <0,001 | <0,001 |
| Frecuencia-Monto | 2 | -0,459 | 0,356 | <0,001 | <0,001 | 15% | 19% |
| 3 | -0,427 | 0,554 | <0,001 | <0,001 |
| 4 | -0,475 | 0,635 | <0,001 | <0,001 |
| 5 | -0,483 | 0,676 | <0,001 | <0,001 |
| 6 | -0,487 | 0,659 | <0,001 | <0,001 |
| 7 | -0,665 | 0,766 | <0,001 | <0,001 |
| 8 | -0,799 | 0,978 | <0,001 | <0,001 |
| 9 | -0,844 | 1,113 | <0,001 | <0,001 |
| Edad\_cat | (25; 27] | -0,430 | 0,307 | <0,001 | <0,001 | 16% | 12% |
| (27; 32] | -0,647 | 0,486 | <0,001 | <0,001 |
| (32; Inf] | -0,790 | 0,599 | <0,001 | <0,001 |
| Flag\_Veh | Si | -0,389 | 0,526 | <0,001 | <0,001 | 6% | 8% |
| Zona | 2 | 0,207 | -0,359 | <0,001 | <0,001 | 12% | 8% |
| 3 | 0,654 | -0,433 | <0,001 | <0,001 |

Tabla 10 Parámetros regresión Logística Multinomial

*En donde el parámetro estimado corresponde al estadístico de media consolidando el resultado de las 100 iteraciones. Dado que la estimación final del parámetro consiste a un promedio, se utiliza el teorema central del límite para estimar el valor-p en base a la distribución normal.*

En la **Tabla 10** se observa que todas las variables resultan significativas bajo un nivel de confianza del 95%. También, se confirma que la variable aporta equitativamente a la estimación del modelo manteniendo todo un peso mayor al 5%.

Los resultados en las 100 iteraciones del modelo y el diccionario de variables seleccionadas pueden ser revisadas en el siguiente documento adjunto.



Por otro lado, la construcción de las regresiones cuantiles se hace en base a la distribución real de ingresos. Es decir, todas las observaciones con una renta menor a 1.639 soles (19.735 cuenta sueldo) son destinadas a la construcción de la regresión cuantil para el segmento 1. Todas las cuentas sueldo con ingresos mayores a 1.639 soles y menores a 3.240 soles (13.095 cuenta sueldo) son destinadas a la construcción del segmento 2. Por último, con las 5.040 cuentas sueldo restantes con un ingreso mayor a 3.240 soles son destinadas a la construcción del segmento 3. Los modelos resultantes para cada segmento son observados en las siguientes tablas:

Tabla 11 Cuantil segmento 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Intervalo** | **Beta** | **Valor-p** | **Peso de Variables** |
| (Intercept) |  | 990 | <0,001 |  |
| CON\_SALDO\_max | (5.565,55 a 19.809,84] | 40 | <0,001 | 22% |
| Mayor a 19.809,84 | 100 | <0,001 |
| TOTAL\_SALDO\_cp\_mean2\_pos | Mayor a 577.77 | 46 | <0,001 | 15% |
| cupo\_tc\_mean\_median | (3.495,75 ; 8.413,92] | 73 | <0,001 | 30% |
| Mayor 8.413,92 | 121 | <0,001 |
| Edad\_cat | (25 ; 27] | 23 | <0,001 | 16% |
| (27 ; 32] | 43 | <0,001 |
| (32, Inf] | 46 | <0,001 |
| Zona | 3 | -43 | <0,001 | 16% |

Tabla 12 Cuantil Segmento 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Intervalo** | **Beta** | **Valor-p** | **Peso de Variables** |
| (Intercept) |  | 2.500 | <0,001 |  |
| CON\_SALDO\_max | Mayor a 19.809,84 | 47 | <0,001 | 28% |
| cupo\_tc\_mean\_median | Mayor a 8.413 | 48 | <0,001 | 30% |
| Zona | 2 | -48 | <0,001 | 42% |
| 3 | -98 |

Tabla 13 Cuantil Segmento 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Intervalo** | **Beta** | **Valor-p** | **Peso de Variables** |
| (Intercept) |  | 4.177 |  |  |
| TOTAL\_SALDO\_cp\_mean2\_pos | (1.328,04571428571; 3.630,73] | 201 | <0,001 | 27% |
| Mayor a 3.630,73 | 501 |
| CON\_TRX\_POR\_UTIL\_max | Mayor a 0.36 | -117 | <0,001 | 9% |
| cupo\_tc\_mean\_median | Mayor a 18.975,01 | 411 | <0,001 | 26% |
| FRECUENCIA-MONTO | 7, 8 o 9 | 119 | <0,001 | 9% |
| Flag\_Veh | Si | 236 | <0,001 | 16% |
| Zona | 2 o 3 | -173 | <0,001 | 12% |

Los resultados detallados de la estructura de las variables y los resultados de las 100 iteraciones realizadas en la validación cruzada pueden ser revisado en los siguientes documentos adjuntos:

# **Desempeño de Modelos**

El principal factor de medición en el desempeño del modelo es en referencia a la sobrestimación, subestimación y estimación correcta. En base a las estrategias internas del banco, se define a un ingreso como sobrestimado cuando la estimación del mismo supera el valor real con un error relativo absoluto mayor al 25%. En cambio, se considerará un ingreso como subestimado si la estimación del mismo se encuentra por debajo del valor real y el error relativo absoluto es mayor a un 25%. Todo ingreso que no cumpla con ninguna de las 2 condiciones anteriores se considerará estimado correctamente. Para evitar altos niveles de sobreendeudamiento, los modelos de ingreso dentro de Banco Falabella Perú tienen por requisito no superar el 25% de ingresos sobrestimados en base a toda la población estudiada.

Otros indicadores que se utilizan en la medición de desempeño del modelo es el error promedio, y el error relativo promedio.

Con el objetivo de medir el correcto desempeño de los modelos, resulta necesario comprender tanto el desempeño conjunto como el individual. En primer lugar, en las siguientes tablas se exponen los resultados obtenidos para la regresión logística multinomial en el proceso de segmentación:

Tabla 14 Desempeño Regresión Logística Multinomial

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Muestra de construcción** | **Muestra de validación** | **Desempeño Final** |
| **Correcto** | 61,00% | 60,84% | 60.97 % |
| **Sobrestimado** | 10,24% | 10,31% | 10.19 % |
| **Subestimado** | 28,76% | 28,84% | 28.84 % |

*Los desempeños expuestos para la muestra de construcción y validación corresponden a las tendencias promedio encontradas para las 100 iteraciones realizadas dentro de la validación cruzada. La columna de desempeño Final hace referencia a los resultados obtenidos posterior estimación final de los parámetros en base a las 100 muestras bootstrap.*

Tabla 15 Desempeño Final Regresión Logística Multinomial

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Predicción** | | |
| **1** | **2** | **3** |
| **Real** | **1** | 44.01% | 7.59 % | 0.51 % |
| **2** | 19.38 % | 13.11 % | 2.08 % |
| **3** | 3.62 % | 5.83 % | 3.85 % |

*Los resultados expuestos en la Tabla 15 corresponden a los segmentos estimados en base a la estimación bootstrap de las 100 iteraciones.*

En este punto, podemos observar que el 60,97% de los ingresos son estimados en la categoría que les corresponde. En cambio, el 10,19% de los ingresos fueron asignados a una categoría con ingresos más altos que los reales, mientras que el 28,84% restante fueron asignados a categorías con ingresos inferiores a los reales.

Condicionado a los resultados anteriores, el desempeño del modelo para cada uno de los cruces observados en la matriz de confusión (**Tabla 15**) es resumido en la siguiente tabla:

Tabla 16 Desempeño Global del Modelo

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Seg. estimado** | **Seg. Real** | **Sobrestimado** | **Subestimado** | **Correcto** | **N** | **N % Total** | **N% Seg. Est.** | **N% Seg. Real** | **Asertividad** | **Error Promedio** | **Error Relativo Promedio** | **Desv. Est. Est.** | **Desv. Est. Real.** |
| 1 | 1 | 4,92% | 18,26% | 76,82% | 16.665 | 65,67% | 84,44% | 84,44% | 86,40% | 176,52 | 14,72% | 46,95 | 240,77 |
| 1 | 2 | 0,00% | 100,00% | 0,00% | 7.341 | 28,93% | 56,06% | 56,06% | 47,24% | 1.252,22 | 52,76% | 48,40 | 432,35 |
| 1 | 3 | 0,00% | 100,00% | 0,00% | 1.371 | 5,40% | 27,20% | 27,20% | 25,35% | 3.275,98 | 74,65% | 46,54 | 952,08 |
| 2 | 1 | 100,00% | 0,00% | 0,00% | 2.876 | 28,61% | 14,57% | 14,57% | 49,12% | 1.265,45 | 113,24% | 38,67 | 255,52 |
| 2 | 2 | 23,77% | 0,04% | 76,19% | 4.965 | 49,39% | 37,92% | 37,92% | 86,37% | 361,36 | 16,56% | 39,31 | 438,20 |
| 2 | 3 | 0,00% | 95,84% | 4,16% | 2.211 | 22,00% | 43,87% | 43,87% | 58,38% | 1.993,31 | 41,62% | 40,84 | 1.066,75 |
| 3 | 1 | 100,00% | 0,00% | 0,00% | 194 | 7,95% | 0,98% | 0,98% | 26,59% | 3.382,23 | 296,14% | 219,87 | 251,75 |
| 3 | 2 | 100,00% | 0,00% | 0,00% | 789 | 32,32% | 6,03% | 6,03% | 54,66% | 2.113,68 | 88,93% | 225,29 | 414,17 |
| 3 | 3 | 15,64% | 20,16% | 64,20% | 1.458 | 59,73% | 28,93% | 28,93% | 82,48% | 1.019,65 | 19,50% | 268,15 | 1.265,08 |
|  |  | 16,07% | 37,42% | 46,51% | 37.870 |  |  |  | 71,01% | 799,50 | 36,73% | 1.025,12 | 1.272,25 |

En la última fila de la Tabla 16 es posible observar el desempeño global del modelo completo. En donde se ejecuta de manera conjunta la regresión logística multinomial y la regresión cuantil por segmento estimado. Se observa una estimación correcta el 46,51% en donde los ingresos sobrestimados no superan el 25% de la población. Por otro lado, cuando el segmento fue estimado correctamente, la estimación también mantiene niveles acordes de sobrestimación.

# ANEXOS

## Anexo (Creación de Variables)

Con el objetivo de construir un conjunto de variables significativas estadísticamente como consistentes a nivel de negocio, en el primer paso se resume la información mensual presente en el sistema financiero por medio de 18 medidas distintas:

1.- Numeró de Visualizaciones en los últimos 12 meses (\_obs).

2.- Numeró de Visualizaciones negativas en los últimos 12 meses (\_neg\_obs).

3.- Numeró de Visualizaciones positivas en los últimos 12 meses (\_pos\_obs).

4.- Promedio en los últimos 12 meses (\_mean).

5.- Logaritmo natural del promedio en los últimos 12 meses (ln\_mean).

6.- Mediana en los últimos 12 meses(\_median).

7.- Mínimo en los últimos 12 meses (\_min).

8.- Máximo en los últimos 12 meses (\_max).

9.- Decrecimiento promedio en los últimos 12 meses (\_cp\_mean\_pos).

10.- Decrecimiento promedio mensual en los últimos 12 meses (\_cp\_mean2\_pos).

11.- Crecimiento promedio en los últimos 12 meses (\_cp\_mean\_neg).

12.- Crecimiento promedio mensual en los últimos 12 meses (\_cp\_mean2\_neg).

13.- Logaritmo natural del decrecimiento promedio mensual en los últimos 12 meses (\_cp\_ln\_mean2\_pos).

14.- Mínimo decrecimiento en los últimos 12 meses (\_cp\_min\_pos).

15.- Mínimo crecimiento en los últimos 12 meses (\_cp\_min\_neg).

16.- Máximo decrecimiento en los últimos 12 meses (\_cp\_max\_pos).

17.- Máximo crecimiento en los últimos 12 meses (\_cp\_max\_neg).

18.- Mediana de los decrecimientos en los últimos 12 meses (cp\_median\_pos).

19.- Mediana de los crecimientos en los últimos 12 meses (cp\_median\_neg).

Un ejemplo de las reglas comentadas en el punto anterior puede observarse con el siguiente caso:





Para comprender los cálculos, se adjunta el Excel con el ejemplo descrito anteriormente.



Para la ejecución automática de estas 18 mediciones en cada una de las variables fue necesario la utilización del programa *RStudio* por medio del algoritmo expuesto a continuación:

### Carga de Paquetes

#####

# rm(list = ls())

library(readr)

library(dplyr)

library(xlsx)

library(tidyr)

library(ggplot2)

library(readxl)

library('gmodels')

library(haven)

library(stringr)

#####

#### Lectura de da

RCC = read\_sav("\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*")

RCC = RCC %>% replace\_na(list(RCC=0))

RCC = RCC %>% mutate(CON\_TRX\_POR\_UTIL = ifelse(cupo\_tc<=10, NA, CON\_TCRC\_SALDO/cupo\_tc),

CON\_TRX\_NOUTIL = ifelse(cupo\_tc<=10, NA, cupo\_tc- CON\_TCRC\_SALDO),

cupo\_tc\_mean = ifelse(cupo\_tc<=10, NA, cupo\_tc/NUM\_TC))

#### Creación de variables

#####

aux\_fun = function(df=RCC, name\_var="TOTAL\_SALDO" ){

sub\_fun = function(var1,var2){

if(sum(!is.na(var1)) == 1){

return(NA)

}else{

return( c(NA ,var2[1:(sum(!is.na(var1))-1)]))

}

}

df %>%dplyr::select(NumDoc, Periodo, var = name\_var) %>% group\_by(NumDoc) %>%

arrange(Periodo)%>% group\_by(NumDoc) %>%

mutate(orden = order(Periodo,decreasing = F),var = var[orden])%>% na.omit() %>% group\_by(NumDoc) %>%

mutate(orden\_desf = sub\_fun(var,orden),var\_desf = sub\_fun(var,var),

pag\_med =var\_desf-var,pag\_med\_pos = ifelse(pag\_med<=0, NA, pag\_med),

pag\_med\_neg = ifelse(pag\_med>=0,NA , pag\_med ), dif\_orden = abs(orden-orden\_desf) )%>%

group\_by(NumDoc) %>% summarise(

######### Comportamiento general

#### Numero de apariciones

pag\_obs = sum(!is.na(var)),

pag\_pos\_obs = sum(!is.na(pag\_med\_pos)),

pag\_neg\_obs = sum(!is.na(pag\_med\_neg)),

#### Media

pag\_mean = mean(var, na.rm = T),

pag\_ln\_mean = ifelse(pag\_mean == 0, NA, log(pag\_mean)),

#### Mediana

pag\_median = median(var, na.rm = T),

#### Minimo

pag\_min =ifelse(pag\_obs==0, NA, min(var, na.rm = T)),

#### Maximo

pag\_max =ifelse(pag\_obs==0, NA, max(var, na.rm = T)),

######### Comportamientros de pago

#### Promedio de los pagos

pag\_cp\_mean\_pos = ifelse(pag\_pos\_obs==0, NA, mean(pag\_med\_pos, na.rm = T)),

pag\_cp\_mean\_neg = ifelse(pag\_neg\_obs==0, NA, mean(pag\_med\_neg, na.rm = T)),

pag\_cp\_mean2\_pos = ifelse(pag\_pos\_obs==0, NA, mean(pag\_med\_pos/dif\_orden, na.rm = T)),

pag\_cp\_ln\_mean2\_pos = ifelse(pag\_cp\_mean2\_pos==0, NA, log(pag\_cp\_mean2\_pos)),

pag\_cp\_mean2\_neg = ifelse(pag\_neg\_obs==0, NA, mean(pag\_med\_neg/dif\_orden, na.rm = T)),

pag\_cp\_ln\_mean2\_neg = ifelse(pag\_cp\_mean2\_neg==0, NA, log(abs(pag\_cp\_mean2\_neg))),

#### Mediana de los pagos

pag\_cp\_median\_pos = median(pag\_med\_pos, na.rm = T),

pag\_cp\_median\_neg = median(pag\_med\_neg, na.rm = T),

#### Pago minimo

pag\_cp\_min\_pos =ifelse(pag\_pos\_obs==0, NA, min(pag\_med\_pos, na.rm = T)),

pag\_cp\_min\_neg =ifelse(pag\_neg\_obs==0, NA, min(pag\_med\_neg, na.rm = T)),

#### Pago Maximo

pag\_cp\_max\_pos =ifelse(pag\_pos\_obs==0, NA, max(pag\_med\_pos, na.rm = T)),

pag\_cp\_max\_neg =ifelse(pag\_neg\_obs==0, NA, max(pag\_med\_pos, na.rm = T))) %>%

rename\_all(funs(str\_replace(., "pag", name\_var))) %>%

return()

}

n\_v = c("cupo\_tc\_mean")

var\_rcc = c("RCC","cupo\_tc\_mean","cupo\_tc","MAX\_CLASIFICACION","NUM\_ENTIDADES","NUM\_ENTIDADES\_SALDO","NUM\_TC","NUM\_TC\_SALDO","NUM\_PREST","NUM\_VEH","NUM\_HIPO",

"TOTAL\_SALDO","NEG\_SALDO","MIC\_SALDO","CON\_SALDO","CON\_TCRC\_SALDO","CON\_TCRC\_COMPRAS","CON\_TCRC\_DISP\_EFEC",

"CON\_TCRC\_OTROS","CON\_TCRC\_COMPRAS\_VIG","CON\_CCRV\_SALDO","CON\_CCRV\_LINEA\_NO\_UTIL","CON\_CCNR\_SALDO","CON\_CVEH\_SALDO",

"CON\_CCTE\_SALDO","HIP\_SALDO","INTS\_SALDO","CONT\_SALDO", "CON\_TRX\_POR\_UTIL", "CON\_TRX\_NOUTIL")

##### Ejecución iterativa de la función

RCC\_var1 = tibble()

i = 1

for (n\_v in var\_rcc ){

print(paste0("repet: ",i," - Sys.time: ",Sys.time()))

print(paste0("variables: ", n\_v))

if(i==1){

#n\_v = "CON\_SALDO"

RCC\_var1 = aux\_fun(df = RCC, name\_var = n\_v )

}else{

RCC\_var1 = RCC\_var1 %>%

left\_join(aux\_fun(df = RCC, name\_var = n\_v), by = c("NumDoc"))

}

i = i+1

}

RCC\_var1 %>%

write\_sav("\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*")

## Anexo (Segmentación de Variables)

Para la segmentación de variables se hace uso de los árboles de clasificación con metodología CART desarrollados en *RStudio* por medio de la librería **rpart.**

El código expuesto itera en cada variable del sistema financiero, generando un árbol de clasificación con toda la información observada con categorías de un tamaño mínimo del 5% de la población y un parámetro de complejidad 0,005. El código utilizado para la segmentación de variables del sistema financiero es el desarrollado en el siguiente código:

### Instal librerías

library(dplyr)

library(streamlineR)

library(xlsx)

library(haven)

library(dplyr)

library(tidyr)

### Lectura de datos

RCC\_var1 = read\_sav("\*\*\*\*\*\*\*\*\*\*\*")

INGRESO = read\_sav("\*\*\*\*\*\*\*\*\*\*\*\*\*\*")

df\_aux = INGRESO %>% rename(NumDoc = DNI ) %>% left\_join(RCC\_var1, by = c("NumDoc"))

rcc\_auto = function(set\_var = "RCC\_max",df = df\_aux){

sum\_tab\_full = tibble()

tab\_res\_full = tibble()

Data\_cat = tibble(NumDoc = INGRESO$DNI)

i = 1

for(n\_v in set\_var){

#n\_v = "RCC\_neg\_obs"

print(paste0("repet: ",i," - Sys.time: ",Sys.time()))

print(paste0("variables: ", n\_v))

#n\_v = "RCC\_max"

no\_lost = df %>% dplyr::select(var = n\_v) %>% summarise(a = sum(!is.na(var))) %>% as.numeric()

if(no\_lost==0) next

df\_a = df %>% dplyr::select(NumDoc,INGRESO\_COMPRADO, INGRESO\_REAL ,var = n\_v, RCC\_max) %>%

replace\_na( list("RCC\_max"=0))

lg.bin.tts <- bin.rpart(formula = INGRESO\_REAL ~ var, data = as.data.frame(df\_a),

rcontrol = rpart.control(cp=0.005,minbucket = .05 \* no\_lost))

if(any(lg.bin.tts == "No Bin")){

mod\_lin = lm(INGRESO\_REAL ~ var, data = df\_a)

mod\_lin\_sum = summary(mod\_lin)

tab\_res = tibble(Variable = n\_v, Missing = (nrow(df\_a)-no\_lost)/nrow(df\_a),

R\_cuad\_lin = mod\_lin\_sum$r.squared, R\_cuad\_aju\_lin = mod\_lin\_sum$adj.r.squared,

AIC\_lin = AIC(mod\_lin), BIC\_lin = BIC(mod\_lin))

tab\_res\_full = tab\_res\_full %>% bind\_rows(tab\_res)

i = i+1

next

}

df\_a = df\_a %>% mutate(cat = lg.bin.tts$bins)

df\_a = df\_a %>% mutate(cat = ifelse(RCC\_max==1 & cat == "Missing", "Missing\_var", as.character(cat))) %>%

mutate(cat = ifelse(var %in% c(0), "0", cat )) %>%

mutate(cat =factor(cat,levels = c("0", levels(df\_a$cat), "Missing\_var") ) ) %>%

replace\_na( list("cat"="Missing"))

mod\_lin = lm(INGRESO\_REAL ~ var, data = df\_a)

mod\_lin\_sum = summary(mod\_lin)

mod\_cat = lm(INGRESO\_REAL ~ cat, data = df\_a)

mod\_cat\_sum = summary(mod\_cat)

Data\_cat = df\_a %>% dplyr::select(NumDoc,cat) %>%

mutate(cat = as.character(cat)) %>%

rename\_all(funs(str\_replace(., "cat", n\_v))) %>%

right\_join(Data\_cat, by = c("NumDoc"))

sum\_tab = df\_a %>% group\_by(cat) %>%

summarise(N= n(), N\_pro =n()/nrow(df\_a), N\_prop\_nolost = n()/no\_lost,

Ing\_real\_mean = mean(INGRESO\_REAL, na.rm = T), Ing\_comprado\_mean = mean(INGRESO\_COMPRADO, na.rm = T),

Variable = n\_v) %>% mutate(Coefficient = names(mod\_cat$coefficients), Beta = as.numeric(mod\_cat$coefficients))

sum\_tab\_full = sum\_tab\_full %>% bind\_rows(sum\_tab)

tab\_res = tibble(Variable = n\_v, Categorias = nrow(sum\_tab), Missing = (nrow(df\_a)-no\_lost)/nrow(df\_a),

R\_cuad\_lin = mod\_lin\_sum$r.squared, R\_cuad\_aju\_lin = mod\_lin\_sum$adj.r.squared,

AIC\_lin = AIC(mod\_lin), BIC\_lin = BIC(mod\_lin),

R\_cuad\_cat = mod\_cat\_sum$r.squared, R\_cuad\_aju\_cat = mod\_cat\_sum$adj.r.squared,

AIC\_cat = AIC(mod\_cat), BIC\_cat = BIC(mod\_cat))

tab\_res\_full = tab\_res\_full %>% bind\_rows(tab\_res)

i = i+1

}

return(list(Tab\_Full = sum\_tab\_full, Res\_Full = tab\_res\_full, Data\_cat = Data\_cat))

}

set\_variables =names(RCC\_var1)[!(names(RCC\_var1)%in% c("NumDoc","RCC\_max"))]

result =rcc\_auto(set\_var = set\_variables,df = df\_aux)

carpeta = "\*\*\*\*\*\*\*\*\*”

result$Data\_cat %>%

write\_sav(paste0(carpeta, "RCC\_cat0502.sav"))

salida = "RCC\_VAR\_ANA\_0502.xlsx"

result$Tab\_Full %>%

write.xlsx2(paste0(carpeta, salida),

append = F, sheetName = "Tab\_Full")

result$Res\_Full %>%

write.xlsx2(paste0(carpeta, salida),

append = T, sheetName = "Res\_Full")

## Desempeño Segmentos

El desempeño promedio por segmento para las interacciones bootstrap es resumido en las siguientes tablas:

### Segmento Bajo (Regresión Cuantil)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Construcción** | **Test** | **Total** |
| **Sobrestimado** | 6,40% | 6,43% | 6,38% |
| **Subestimado** | 16,74% | 16,77% | 16,73% |
| **Correctos** | 76,85% | 76,80% | 76,89% |
| **N° Observación** | 13.814 | 5.921 | 19.735 |
| **Asertividad** | 85,92% | 85,89% | 85,91% |
| **Error relativo promedio** | 15,43% | 15,47% | 15,44% |
| **Desv. Est. Estimación** | 80,45 | 80,30 | 80,13 |
|  | 4,22% | 4,06% | 4,20% |

### Segmento Medio (Regresión Cuantil)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Construcción** | **Test** | **Total** |
| **Sobrestimado** | 27,91% | 28,00% | 28,04% |
| **Subestimado** | 0,03% | 0,03% | 0,02% |
| **Correctos** | 72,06% | 71,97% | 71,94% |
| **N° Observación** | 9.166 | 3.929 | 13.095 |
| **Asertividad** | 85,71% | 85,69% | 85,71% |
| **Error relativo promedio** | 17,62% | 17,66% | 17,63% |
| **Desv. Est. Estimación** | 43,81 | 43,78 | 43,66 |
|  | 0,35% | 0,37% | 0,36% |

### Segmento Alto (Regresión Cuantil)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Construcción** | **Test** | **Total** |
| **Sobrestimado** | 7,43% | 7,48% | 7,42% |
| **Subestimado** | 15,54% | 15,56% | 15,75% |
| **Correctos** | 77,03% | 76,96% | 76,83% |
| **N° Observación** | 3.528 | 1.512 | 5.040 |
| **Asertividad** | 84,26% | 84,21% | 84,25% |
| **Error relativo promedio** | 17,04% | 17,09% | 17,05% |
| **Desv. Est. Estimación** | 343,85 | 343,65 | 342,62 |
|  | 4,09% | 3,63% | 4,02% |

En donde el asertividad es interpretado como la similitud promedio que posee el ingreso real en comparación al estimado, resumiendo su cálculo a:

Por otro lado, hace referencia a una medida de ajuste lineal, en donde se compara el ajuste del modelo sobre el ajuste del modelo nulo en base a la función de pérdida definida para el percentil estimado.

Sea la función de perdida definida por , entonces la métrica para los modelos de regresión cuantil a la media queda determinada por:

## Anexo Construcción de variables seleccionadas

A continuación, se exponen las fórmulas y la estructura de las variables finales seleccionadas por el modelo.

### Saldo máximo en crédito de consumo de los últimos 12 meses (CON\_SALDO\_max)

La variable tiene por objetivo rescatar el saldo máximo en créditos de consumo de los últimos 12 meses observados. Sean CON\_SALDO\_i la observación del saldo en créditos de consumo de una persona para el mes i. Entonces:

En caso de que la variable CON\_SALDO\_i presente valores perdidos, estos deberán ser remplazados por “0” haciendo referencia a que el cliente no posee créditos de consumo.

Como se comenta en los apartados metodológicos, la variable cambia su segmentación dependiendo del modelo en el que sea utilizado.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 1** | **Regresión Cuantil Segmento 2** |
| CON\_SALDO\_max | <= 5565.55-0-Miss | <= 5565.55-0-Miss | <= 19809.84 – Miss |
| 5565.55 < · <= 19809.84 | 5565.55 < · <= 19809.84 |
| 19809.84 < · <= 43697.65 | >19.809,84 | >19.809,84 |
| > 43697.65 |

### Reducción promedio mensual del saldo total en el sistema financiero. ( TOTAL\_SALDO\_cp\_mean2\_pos)

Esta variable tiene por objetivo representar la capacidad de pago que posee una persona en base a la medición de los decrecimientos mensuales vistos para su saldo total en el sistema financiero. En caso de que una persona tenga meses no reportados y es posible observar un decrecimiento de su deuda en estos intervalos observados, la magnitud del decrecimiento es dividida por la cantidad de meses entre las observaciones.

Sea **TOTAL\_SALDO\_observados** el vector que contiene únicamente las mediciones del saldo total en RCC observadas para último año estudiado, entonces la reducción promedio mensual del saldo total en el sistema financiero podría definirse de la siguiente manera:

Debido a que esta medición se basa en solo la información observada, todo cliente que no tenga información de TOTAL\_SALDO o no se hayan observado decrecimientos en los últimos 12 meses estudiados, será considerado como perdido y agrupado en la categoría con menores ingresos. La ejemplificación en la creación de esta variable puede ser observada en el Anexo (Creación de Variables).

Las categorías definitivas de esta variable según modelo en el cual se utiliza fueron:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 1** | **Regresión Cuantil Segmento 3** |
| TOTAL\_SALDO\_cp\_  mean2\_pos | <= 577,77-Miss | <= 577,77-Miss | <=1.328,04571428571 - Miss |
| 577,77 < · <= 1.328,04571428571 | >577,77 |
| 1.328,04571428571 < · <= 3.630,73 | 1.328,04571428571 < · <= 3.630,73 |
| > 3.630,73 | > 3.630,73 |

### Porcentaje de Utilización Máxima de Línea de crédito en los últimos 12 meses (CON\_TRX\_POR\_UTIL\_max)

El porcentaje de Utilización es calculado en base al saldo utilizado en tarjeta de crédito (CON\_TCRC\_SALDO) dividido por la Línea total disponible (CON\_TCRC\_LINEA\_TC). En caso de que la línea disponible en tarjetas de crédito sea menor a 10 soles, se considera a esta persona como si no tuviera este producto y se considera un *missing* agrupándolo al primer segmento. Sea entonces CON\_TCRC\_SALDO\_i el saldo utilizado en tarjeta de crédito el mes i y CON\_TCRC\_LINEA\_TC\_i el monto en línea de crédito disponible en el mes i, la utilización máxima de los últimos 12 meses puede ser calculada como:

Si por algún motivo, el uso de la tarjeta de crédito excede la línea disponible, este valor es truncado en el 100% considerando al cliente como si haya tenido una utilización máxima.

Los segmentos de esta variable según el modelo al que se incorpora quedan definidos por:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 3** |
| CON\_TRX\_POR\_UTIL\_max | Miss - 0 | <=0.365 - Miss |
| 0 <.<= 0.365 |
| 0.365<.<=0.694 | >0.365 |
| >0.694 |

### Mediana de los últimos 12 meses de la línea en tarjeta de crédito de la media por tarjeta en el sistema financiero (cupo\_tc\_mean\_median).

Esta variable representa la mediana de los últimos 12 meses observados para la línea de crédito de la media por tarjeta que posee una persona. El objetivo es incorporar el poder adquisitivo que posee una persona en base a la línea que posee el cliente independiente del número de tarjetas que este tenga. En el caso de que una persona que no tenga Línea de crédito, o ésta sea inferior a 10 soles fue considerada como perdida y agrupada en la categoría más conservadora.

Sea CON\_TCRC\_LINEA\_TC\_i el monto de línea de las tarjetas de crédito y NUM\_TC\_i el Número de Tarjetas reportadas en el mes i. La medición de esta variable puede ser resumida como.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 1** | **Regresión Cuantil Segmento 2** | **Regresión Cuantil Segmento 3** |
| cupo\_tc\_mean\_  median | <= 3495.75-Miss | <= 3495.75-Miss | <= 8.413,92-Miss | <= 8.413,92-Miss |
| 3495.75 < · <= 8413.92 | (3.495,75 ; 8.413,92] |
| 8413.92 < · <= 18975.01 | Mayor 8.413,92 | Mayor a 8.413 |
| > 18975.01 | > 18975.01 |

### Capacidad de compra dentro del retail (Frecuencia - Monto)

En primer lugar, el importe y el número de compras fueron considerados a nivel consolidado sin hacer distinción la o las tiendas en las cuales el cliente ha realizado la compra. Esto quiere decir que:

* IMPORTE\_TO\_i, es el importe en el mes i desembolsado por un cliente en las tiendas de Tottus.
* IMPORTE\_SA\_i, es el importe en el mes i desembolsado por un cliente en las tiendas de Saga Falabella.
* IMPORTE\_MS\_i, es el importe en el mes i desembolsado por un cliente en las tiendas Maestro.
* IMPORTE\_SO\_i es el importe en el mes i desembolsado por un cliente en las tiendas Sodimac.

Del mismo modo, sea:

* N\_COMP\_TO\_i, el número de veces (días) en el mes i que un cliente ha realizado transacciones en las tiendas de Tottus.
* N\_COMP\_SA\_i, el número de veces (días) en el mes i que un cliente ha realizado transacciones en las tiendas de Saga Falabella.
* N\_COMP\_MS\_i, el número de veces (días) en el mes i que un cliente ha realizado transacciones en las tiendas Maestro.
* N\_COMP\_SO\_i, el número de veces (días) en el mes i que un cliente ha realizado transacciones en las tiendas Sodimac.

El proceso de segmentación sigue los siguientes pasos: tanto la variable de importe total como el número de compras realizadas es segmentada en 9 categorías por medio de deciles, en donde la primera categoría se ve concentrada por todas las personas que no realizan transacciones dentro del grupo.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Importe total realizado en los últimos 12 meses de observación | | | Número de compras totales realizadas en los últimos 12 meses de observación | | |
| Segmento | Número de Casos | Porcentaje en la población | Segmento | Número de Casos | Porcentaje en la población |
| (-1,0] | 13.510 | 27% | (-1,0] | 13.510 | 27% |
| (0,28.7] | 1.425 | 9% | (0,1] | 4.533 | 3% |
| (28.7,204] | 4.976 | 5% | (1,2] | 2.717 | 10% |
| (204,603] | 4.978 | 10% | (2,5] | 4.911 | 10% |
| (603,1.270] | 4.977 | 10% | (5,10] | 4.948 | 10% |
| (1.270,2.290] | 4.978 | 10% | (10,17] | 4.576 | 10% |
| (2.290,3.890] | 4.977 | 10% | (17,28] | 4.627 | 10% |
| (3.890,7.020] | 4.978 | 10% | (28,50] | 5.073 | 10% |
| (7.020,...] | 4.978 | 10% | (50,…] | 4.882 | 10% |

Teniendo cada variable segmentada en 9 categorías, se procede a unificar los importes y los números de compra en una sola variable de frecuencia y monto, la cual hace referencia a la categoría mínima a la que pertenece el cliente entre el importe realizado y el número de veces que ha comprado dentro del grupo.

Sean **IMPORTE\_TOTAL\_ORD** y **N\_COMP\_TOTAL\_ORD** variables ordinales del 1 al 9 según el segmento que les corresponda, la variable **Frecuencia-Monto** queda definida como el mínimo entre estas 2 segmentaciones.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Número de Compras | | | | | | | | | Variable Mixta Frecuencia Monto |
|  |  | (-1,0] | (0,1] | (1,2] | (2,5] | (5,10] | (10,17] | (17,28] | (28,50] | (50,…] |  |
| Monto Total Importado | (-1,0] | 27,14% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 1 |
| (0,28.7] | 0,00% | 2,55% | 0,27% | 0,04% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 2 |
| (28.7,204] | 0,00% | 4,61% | 2,67% | 2,29% | 0,38% | 0,05% | 0,00% | 0,00% | 0,00% | 3 |
| (204,603] | 0,00% | 1,42% | 1,74% | 3,86% | 2,22% | 0,64% | 0,10% | 0,02% | 0,00% | 4 |
| (603,1.270] | 0,00% | 0,40% | 0,57% | 2,36% | 3,49% | 2,09% | 0,82% | 0,24% | 0,03% | 5 |
| (1.270,2.290] | 0,00% | 0,10% | 0,14% | 0,90% | 2,36% | 2,98% | 2,26% | 1,13% | 0,14% | 6 |
| (2.290,3.890] | 0,00% | 0,03% | 0,06% | 0,32% | 1,03% | 2,14% | 3,13% | 2,52% | 0,78% | 7 |
| (3.890,7.020] | 0,00% | 0,00% | 0,01% | 0,09% | 0,38% | 1,06% | 2,29% | 3,71% | 2,46% | 8 |
| (7.020,...] | 0,00% | 0,00% | 0,00% | 0,02% | 0,08% | 0,23% | 0,69% | 2,58% | 6,40% | 9 |
| Variable Mixta Frecuencia Monto | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |

Al igual que todas las variables seleccionadas para el modelo, la composición de las categorías puede variar dependiendo del modelo al cual sea incorporando la variable, por lo que se definen las siguientes segmentaciones según modelo.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 1** |
| Frecuencia - Monto | 1 | 1, 2, 3, 4, 5, 6 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 | 7, 8 o 9 |
| 8 |
| 9 |

### Edad segmentada (Edad\_cat)

La edad es calculada utilizando la fecha de nacimiento disponible en las fuentes de información sociodemográfica. Ésta variable es calculada en base a la fecha de referencia 2018-12-31, y utilizando únicamente la unidad de años (en caso de calcular edades con decimales, el resultado debe se debe truncar). En caso de no poseer información, ésta es asignada a la categoría más conservadora.

La variable es segmentada en los siguientes tramos:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 1** |
| Edad | <= 25 - Miss | <= 25 - Miss |
| 25 < · <= 27 | 25 < · <= 27 |
| 27 < · <= 32 | 27 < · <= 32 |
| >32 | >32 |

### Posesión de Vehículo (Flag\_Veh)

Esta variable es extraída de la información sociodemográfica y es generada en base a que el cliente reporte información referente a vehículos a su nombre. En caso de no poseer esta variable se supone que la persona no tiene vehículo.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil Segmento 3** |
| Flag\_Veh | NO | NO |
| SI | SI |

### Zona Geográfica

Esta variable se basa en el resumen de la información de Ubigeo según zona geográfica y niveles de ingreso, incorporando toda información en el segmento 2 (segmento medio). Las segmentaciones realizadas según Ubigeo son expuestas en el siguiente documento adjunto:



Según sea el modelo a realizar, las variables poseen la siguiente segmentación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Segmentos Regresión Multinomial** | **Regresión Cuantil segmento 1** | **Regresión Cuantil segmento 2** | **Regresión Cuantil segmento 3** |
| Zona | 1 | 1-2 | 1 | 1 |
| 2 | 2 | 2-3 |
| 3 | 3 | 3 |

# Referencias Packages

1. quantreg: <https://cran.r-project.org/web/packages/quantreg/quantreg.pdf>
2. nnet: <https://cran.r-project.org/web/packages/nnet/nnet.pdf>
3. rpart: <https://cran.r-project.org/web/packages/rpart/rpart.pdf>
4. cluster: <https://cran.r-project.org/web/packages/cluster/cluster.pdf>

# Referencias

Roger Koenker & José A. F. Machado (1999) Goodness of Fit and Related Inference Processes for Quantile Regression, Journal of the American Statistical Association, 94:448, 1296-1310.

Pando, Valentín & San Martín, Roberto. (2004). Regresión logística multinomial. Cuadernos de la Sociedad Española de Ciencias Forestales, ISSN 1575-2410, Nº. 18.

B. Krishnapuram, L. Carin, M. A. T. Figueiredo and A. J. Hartemink (2005), "Sparse multinomial logistic regression: fast algorithms and generalization bounds," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 957-968.

Hartigan, J., & Wong, M. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* *28*(1), 100-108.