# UNIVERSIDAD NACIONAL DE INGENIERIA

**FACULTAD DE INGENIERIA ECONOMICA, ESTADISTICA Y CIENCIASSOCIALES**

# TRABAJO DE SUFICIENCIA PROFESIONAL

## APLICACIÓN DE LA REGRESIÓN LOGÍSTICA PARA LA CLASIFICACIÓN DE CLIENTES CON CREDITO CONSUMO DEL SECTOR MAGISTERIAL


**ELABORADO POR:**

**GLORIA PRISCILA ROSAS VIDAL**


**Para obtener el Título Profesional de Ingeniero Estadístico**


**ASESOR:**

**MAG. CIRILO ALVAREZ ROJAS**


**LIMA- PERU**

**2016**

# Contenido

**RESUMEN** ... 3

**INTRODUCCIÓN** ... 4

**CAPÍTULO I: REALIDAD PROBLEMÁTICA** ... 6

**1.1 Descripción del problema de la investigación** ... **6**

**1.2 Formulación del Problema General y Específico** ... **7**

**1.2.1 Problema General** ... 7

**1.2.2 Problema Especifico** ... 7

**1.3 Formulación del Objetivo General y Específico** ... **7**

**1.3.1 Objetivo General** ... 7

**1.3.2 Objetivo Especifico** ... 7

**1.4 Justificación** ... **7**

**1.5 Delimitación del estudio** ... **8**

**1.6 Limitación del estudio** ... **8**

**CAPÍTULO II: MARCO TEÓRICO** ... 9

**2.1 Antecedentes de la investigación** ... **9**

**2.2 Bases Teóricas: Generales y Especializadas** ... **11**

**2.3 Formulación de Hipótesis** ... **46**

**2.3.1 Hipótesis Principal** ... 46

**2.3.2 Hipótesis Secundaria** ... 46

**2.4 Variables y Operacionalización** ... **46**

**2.5 Marco Legal** ... **47**

**2.6 Glosario de términos** ... **50**

**CAPÍTULO III: MARCO METODOLÓGICO** ... 52

**3.1 Tipo, Nivel y Diseño de la investigación** ... **52**

**3.2 Población y Muestra** ... **52**

**CAPÍTULO IV: RESULTADOS Y DISCUSIÓN** ... 53

**4.1 Resultados Preliminares** ... **53**

**4.1.1 Análisis Bivariado** ... **53**

**4.2 Construcción del Modelo Logístico** ... **71**

**4.3 Validación del Modelo** ... **76**

**4.4 Discusión de los resultados** ... **86**

**Bibliografía** ... **88**

**4.4.1 Análisis Univariado** ... **90**

**4.4.2 Análisis Bivariado** ... **97**

## RESUMEN

El objetivo fundamental de la presente investigación, fue diseñar un modelo de *CreditScoring*para calcular el riesgo de crédito de un solicitante al ser evaluadoen una Entidad Financiera que se dedica al otorgamiento de créditos consumo no revolvente a docentes que pertenecen al Sector Magisterial. Se plantea la hipótesis de que la probabilidad de incumplimiento de pago de un solicitante en el futuro está determinada por sus antecedentes crediticios en el sistema bancario como son el monto de deuda, número de entidades, calificación crediticia y datos socios demográficos. En la presente investigación, el modelo de creditscoring emplea específicamentelaregresión logística como modelo estadísticoel cual permite calcular la probabilidad de incumplimiento de pago que tiene un cliente sabiendo previamente los valores de las variables independientes.Entre las ventajas que tiene la aplicación del modelo de regresión logística tenemos que no exige la normalidad de la distribución de las variables. En este trabajo se busca conocer además que variables están más correlacionadas con la probabilidad de incumplimiento de pago del cliente del Sector Magisterial.



Palabras claves.- riesgo de crédito, modelo de creditscoring, regresión logística, probabilidad de incumplimiento de pago, créditos de consumo no revolvente.

## INTRODUCCIÓN

En el marco de la competitividad en el que actúan las empresas, se tiene que considerar su capacidad ya no solamente para reaccionar; sino que desarrollen su capacidad para anticipar y responder a estos ambientes tan dinámicos e inciertos y a la vez más exigentes. Las empresas financieras ponen a prueba las competencias de sus directivos y de sus colaboradores en el campo de la creatividad e innovación para crear modelos más modernos para medir el riesgo de crédito como son los modelos decreditscoringque califica a los clientes según su perfil de riesgo y que permiten incrementar la rentabilidad de la empresa minimizando las pérdidas de incumplimiento de pagos (Default).

El mercado crediticio y el desarrollo de los mercados financieros, implica que las entidades financieras deban desarrollar técnicas de medición de riesgo de crédito cada vez más sofisticadas, las cuales tienen por objetivo asignar eficientemente los créditos mediante la adecuada calificación de clientes que puedan cumplir con sus responsabilidades contractuales de pago de sus créditos y los intereses del crédito hasta la madurez del mismo de un modo en que se reduzcan los tiempos de otorgamiento.

La presente investigación resulta importante, ya que permitirá mejorar el nivel de créditos en una Entidad Financiera dedicada al mercado crediticio del Sector Magisterial que se encuentra bajo el régimen de dos normativas las cuales han disminuido la Capacidad de Endeudamiento del docente para comprometerse con Instituciones Financieras mediante créditos de convenio de descuento por planilla. Con este trabajo, se podrá mejorar el proceso de admisión crediticia e incluir un elemento técnico estadístico que permita hacer más eficiente el otorgamiento de créditos.

La Entidad Financiera actualmente tiene políticas de admisión de clientes basadas en Pautas de Créditos establecidas bajo criterio experto los cuales no cuentan con sustento estadístico. Es por ello que se requiere implementar un modelo de creditscoring en la Entidad Financiera utilizando la técnica de Regresión Logística mediante el cálculo de la probabilidad de incumplimiento de pago y que además presente la menor tasa de error en la clasificación de los clientes. La técnica de Regresión Logística es la técnica elegida debido a las siguientes ventajas: La variable dependiente o respuesta presenta dos categorías que representan la ocurrencia y no ocurrencia del acontecimiento definido en el estudio(en el presente estudio son cliente malo y cliente bueno), codificándose con los valores uno y cero, respectivamente; con respecto a las variables independientes o explicativas, no se establece ninguna restricción, pudiendo ser cuantitativas y categóricas. El modelo de regresión logística expresa la variable respuesta en términos de probabilidad, utilizando la función logística para estimar la probabilidad de que ocurra el acontecimiento dados determinados valores de las variables independientes. Ello permite que la interpretación de sus indicadores sea más fácil de explicar.

La población objetivo está comprendida por la base de datos de la cartera de créditos de la Entidad Financiera que contiene la información de las variables demográficas, variables socioeconómicas, variables crediticias externas de los clientes que solicitaron crédito durante el año 2013-2014.

Esta tesis fue estructurada con una serie secuencial de Capítulos:

En el **primer capítulo** que se titula Realidad Problemática, abarcó la descripción del contexto del problema, formulación del problema general y los problemas específicos, formulación del objetivo general y objetivos específicos, justificación, importancia, alcance del estudio y limitaciones.

En el **segundo capítulo**que se titula Marco Teórico,se describirán losantecedentes de la investigación, las bases teóricas generales y especializadas, Formulación de la Hipótesis, Variables y Operacionalización, Marco Legal y Glosario de Términos.

En el **tercer capítulo**que se titula Marco Metodológico, se detalla el tipo, el Nivel y el diseño de la investigación, Población, Muestra, Técnicas e Instrumentos de recolección de datos además de las Técnicas de procesamiento de datos.

En el **cuarto capítulo**, que se titula Resultados y discusión, abarca los resultados preliminares, implementación de las pruebas de hipótesis y la discusión de Resultados.

Finalmente, se exponen las conclusiones y recomendaciones del Informe de Suficiencia.

# CAPÍTULO I: REALIDAD PROBLEMÁTICA

## 1.1 Descripción del problema de la investigación

La Entidad Financiera del estudio brinda distintos servicios a los docentes del Magisterio Peruano como son: previsión social, inmobiliaria, hotelería, programas educativos. Siendo su principal core de negocio: las campañas de crédito que se ofrecen con el objetivo de mejorar la calidad de vida de los docentes del Estado Peruano.

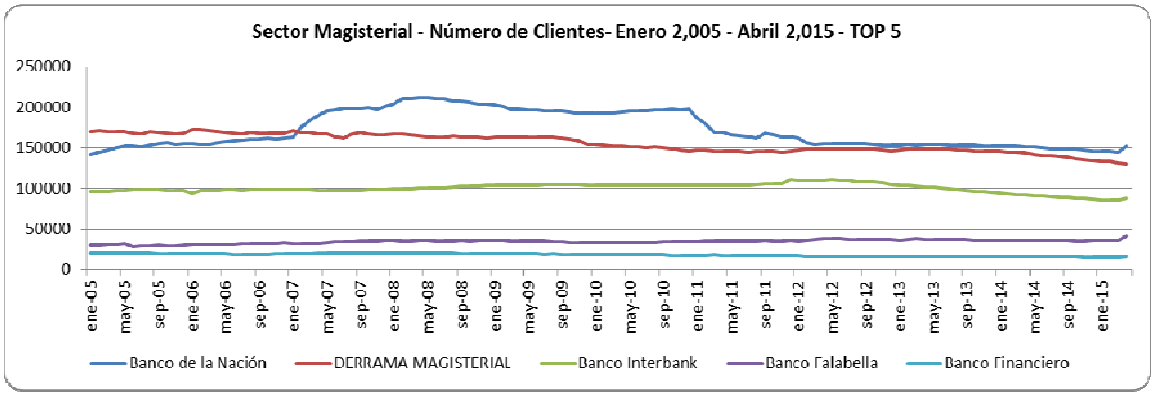
Actualmente en el Marco Legal que regula el procedimiento de cobranza de las cuotas de los créditos mediante convenio de descuento por planilla de los docentes del Sector Magisterial está compuesto por dos leyes: Ley N° 30114 "Ley de Presupuesto del Sector Público" y DS N° 010-2014 las cuales han disminuido el nivel de descuento permitido en los descuentos por planilla que efectúa el MINEDU de 75% al 50%.

Además en la actualidad el área comercial de la Entidad Financiera del estudio tiene políticas de admisión de clientes basadas en Pautas de Créditos establecidas bajo criterio experto los cuales no cuentan con sustento estadístico.Debido a ello se hace necesario determinar un modelo de regresión logística que permita clasificar a los clientes y pronosticar la probabilidad de incumplimiento de pago de los clientes, el objetivo es tratar de ayudar al área de créditos a que incremente el nivel de colocaciones, que se mejoren los tiempos de otorgamiento de créditos y la priorización de clientes clasificados como buenos. Además de ese modo se podrá mitigar el riesgo de admitir clientes morosos.

A continuación se analiza el número de Clientes que comprenden el Sector Magisterial en el cual se reporta que el número de clientes viene disminuyendo a partir del año 2011 ello nos indica que para frenar dicho decrecimiento, por lo tanto se deben mejorar y optimizar el proceso de admisión de clientes tratando de captar y retener a los clientes con los mejores perfiles.

Gráfico N° 1.1

Número de Clientes del TOP 5 en el Sector Magisterial




FUENTE: RCC a Abril 2015

ELABORACION: Propia

## 1.2 Formulación del Problema General y Específico

### 1.2.1 Problema General

No existe un modelo estadístico que permita clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria según su probabilidad de impago.

### 1.2.2 Problema Especifico

Hasta el momento no se ha utilizado un modelo estadístico para determinar que variables son significativas para clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria según su probabilidad de impago

## 1.3 Formulación del Objetivo General y Específico

### 1.3.1 Objetivo General

Elaborar un modelo de regresión logística que permita clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria, según su probabilidad de impago.

### 1.3.2 Objetivo Especifico

Determinar que variables son significativas en el modelo de regresión logística para clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria según su probabilidad de impago.

## 1.4 Justificación

El estudio se justifica debido a que existe la necesidad de pronosticar la probabilidad de incumplimiento de pago de una cartera de clientes comprendida por créditos consumo con el fin de reducir las pérdidas esperadas al otorgar crédito. Asimismo, la aplicabilidad de la investigación está garantizada ya que la Entidad Financiera no dispone de un scoring para su cartera de consumo.

Por lo tanto la realización del score crediticio tendría los siguientes beneficios para la Entidad Financiera debido a que: disminuye el costo de las evaluaciones crediticias y reducen los tiempos de la admisión de créditos, se estandarizan las evaluaciones crediticias, facilitan la implantación de estrategias de negocio diferenciadas de acuerdo a las variables identificadas como significativas. Lo más importante se puede evaluar la efectividad del modelo antes de implementarlo.

El uso del modelo de regresión logística se justifica debido a que presenta las siguientes bondades:La variable dependiente tiene dos categorías que representan la ocurrencia o no de un acontecimiento (codificado por 1 y 0), las variables independientes no tienen restricción pudiendo ser cuantitativas (continuas o discretas) y categóricas. El modelo de regresión logística expresa la variable dependiente en términos de probabilidad

utilizando la función logística que estima la probabilidad de ocurrencia de un evento dados ciertos valores de las variables independientes. Ello hace que la interpretación del modelo sea muy fácil de explicar. Además se puede medir la capacidad predictiva del modelo mediante la comparación de los pronósticos hallados por la regresión logística y los datos observados.

### 1.5 Delimitación del estudio

El estudio se delimita al análisis de los clientes a los cuales se les otorgó crédito durante el año 2,013 al 2,014 mediante el cual se recopilaran las variables demográficas y crediticias principales y definir su probabilidad de incumplimiento de pago conforme a su comportamiento de pago durante el pago total de su crédito.

### 1.6 Limitación del estudio

Para el presente estudio se presentaron las siguientes limitaciones:

- Los registros de la base de datos disponible solo indican los datos personales como son: nombres, edad, datos de ubicabilidad, los ingresos económicos, la UGEL a la que pertenece el docente, lugar de trabajo, estado civil y número de dependientes sin embargo no se cuenta con datos como son: ingresos adicionales, grado de instrucción pudiendo estas variables estar relacionadas con la probabilidad de impago.
- La información crediticia del cliente se reporta con un mes de retraso debido a demoras en la carga de datos de las Centrales de Riesgo externas.

# CAPÍTULO II: MARCO TEÓRICO

## 2.1 Antecedentes de la investigación

D. García Pérez de Lema, A. Arqués Pérez y A. Calvo-Flores Segura (1,995) realizaron un estudio entre la relación de la morosidad de los créditos bancarios de los clientes y los ratios económicos-financieros utilizando la técnica estadística multivariada del análisis discriminante. Mediante este método obtenemos una función lineal de los ratios cuyo objetivo es clasificar a cada observación (una empresa, caracterizada por sus ratios) en uno de los dos grupos: moroso o normal, de acuerdo a la puntuación otorgada a la empresa por la función discriminante.

Los modelos presentados en el estudio provienen de aplicar el análisis discriminante sobre un conjunto de variables que no presentan intensas correlaciones, y coherentes con los principios económico-financieros en relación a la situación de morosidad. Las conclusiones del estudio fueron las siguientes: se construyó una primera función discriminante no sectorizada en la que se obtuvo un nivel de acierto superior al 76%. Luego se separó a las empresas en dos sectores (las constructoras y no constructoras), los resultados mejoran considerablemente dado que los porcentajes de aciertos son 83% para las empresas constructoras y 89% en las no constructoras. Por lo que resulta conveniente realizar el análisis discriminante a la data sectorizada.

Geraldine Judith Vigo Chacón (2,010)el estudiorealiza la comparación de dos métodos clásicos de clasificación: Análisis de Regresión Logística y Árboles de Clasificación, con el método de Redes Neuronales en función al poder de clasificación y predicción de los modelos obtenidos. En el estudio se establecen las ventajas y desventajas en el empleo de cada método.Los datos que se usaron el estudio corresponden a un grupo de personas que solicitaron un préstamo en un Banco Alemán. El objetivo de este trabajo es evaluar la clasificación de un cliente en base al préstamo que se le otorgó en el banco y determinar si un nuevo cliente que solicité un préstamo es un buen o mal pagador.

Las conclusiones del trabajo son: No existe diferencia en el porcentaje de error de entrenamiento entre la Regresión Logística y el árbol de clasificación CART; sin embargo en el caso de la Red Neuronal se obtiene un 84.18% de buena clasificación y 74.32% de buena predicción. El mayor error de en la clasificación y predicción de la Regresión Logística y el árbol de Clasificación CART se debe a que estos modelos son sensibles a los valores influyentes y la Red Neuronal no se afecta de los valores influyentes.

M. Jesús Mures Quintana, Ana García Gallego, M. Eva Vallejo Pascual (2,005) realizaron un estudio para aplicar dos técnicas multivariadas (análisis discriminante y regresión logística) a una muestra de clientes de Entidades Financieras de Castilla y León con el objetivo de clasificar a los clientes en grupos de riesgo. Adicionalmente se determina los factores que influyen en el comportamiento de pago del cliente.

Las conclusiones del análisis son: Ambos modelos son significativo y tienen una alta capacidad predictiva en los clientes no morosos con tasas de acierto de 100% en el análisis discriminante y 98.1% en la regresión logit. Para los clientes no morosos la tasa de acierto 88.9% en el análisis discriminante y 94.4 % en la regresión logística. En ambos casos los modelos tiene una tasa de acierto de 97.1% global. En ambo modelos han resultado significativas la de numero de impagos y a la residencia del cliente, en la regresión logística se incluyen otras dos, quehacen referencia al destino de la financiación solicitada y en la función discriminante se refieren a laduración de los retrasos en el pago producidos con anterioridad, al estadocivil del cliente y si éste ha aportado garantía real, ya sea de tipohipotecario u otra.

## 2.2 Bases Teóricas: Generales y Especializadas

### Clasificación de las técnicas de análisis multivariante

Una vez definido el objetivo del estudio, el investigador tiene que afrontar el problema de seleccionar la técnica estadística adecuada que cumpla con el objetivo del estudio y resuelva las preguntas planteadas. Es por ello que se hace necesario conocer la clasificación de las técnicas de análisis multivariantey las diferencias que existen entre ellas para elegir correctamente que técnica a utilizar de acuerdo al tipo de datos que se tiene y al objetivo del estudio. Las técnicas de Análisis Multivariante de datos se clasifican en tres grupos:

Tabla N° 2.1

Clasificación de Técnicas de Análisis Multivariado

<table>
<tr><td colspan="3">Métodos de Interdependencia: También llamado técnicas multivariantes descriptivas, se utiliza cuando todas las variables tienen la misma importancia, es decir ninguna destaca como dependiente principal o depende de las demás variables.</td></tr>
<tr><td>Objetivo</td><td>Variables</td><td>Técnica</td></tr>
<tr><td rowspan="3">Métodos para reducir la dimensión de un conjunto de variables a través de un conjunto de variables ficticias que son combinación de las variables observadas.</td><td rowspan="2">Métricas</td><td>Análisis de componentes principales</td></tr>
<tr><td>Análisis Factorial</td></tr>
<tr><td>No Métricas</td><td>Análisis de correspondencias</td></tr>
<tr><td rowspan="2">Métodos para clasificar clientes en grupos con cierta homogeneidad</td><td>Métricas y no Métricas</td><td>Análisis Cluster o Análisis de Conglomerados</td></tr>
<tr><td>Métricas y no Métricas</td><td>Escalamiento Multidimensional</td></tr>
<tr><td colspan="3">Métodos de Dependencia: También llamado técnicas multivariantes explicativas, se utiliza cuando por lo menos alguna variable destaca como dependiente principal o depende de las demás variables.</td></tr>
<tr><td>Técnica</td><td>Variable Dependiente</td><td>Variables Independientes</td></tr>
<tr><td>Regresión Múltiple</td><td>Métrica</td><td>Métricas</td></tr>
<tr><td>Regresión Múltiple con variables ficticias</td><td>Métrica</td><td>Métricas y No Métricas</td></tr>
<tr><td>Análisis de Correlación Canónico</td><td>Métricas y No Métricas</td><td>Métricas y No Métricas</td></tr>
<tr><td>Regresión Logística</td><td>No Métrica y dicotómica.</td><td>Métricas y No Métricas</td></tr>
<tr><td>Análisis Discriminante</td><td>No Métrica y Categórica</td><td>Métricas</td></tr>
<tr><td>ANOVA y MANOVA</td><td>Métrica (métricas)</td><td>No Métricas</td></tr>
</table>

| | | |
|---|---|---|
| **ANCOVA y MANCOVA** | Métrica (métricas) | Métricas y No Métricas |
| **Análisis Conjunto** | Métrica o No Métrica | No Métricas |
| **Segmentación** | Métrica o No Métrica | No Métricas |

Elaboración: Propia

Clasificación de las técnicas de Minería de Datos

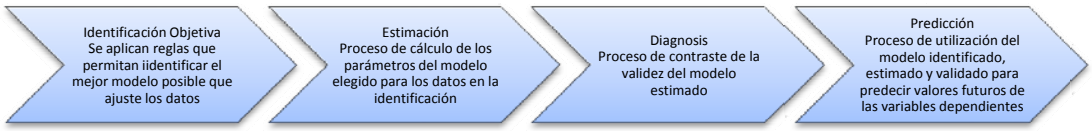
El análisis multivariante se ha transformado en Data Mining o Minería de Datos debido a la necesidad de trabajar con grandes volúmenes de datos. La Minería de Datos tiene como objetivo encontrar relaciones, perfiles y tendencias utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas estadísticas avanzadas de análisis multivariante de datos.

La clasificación de las técnicas estadísticas de análisis multivariado y la clasificación de las técnicas de Data Mining son similares. Es decir, las técnicas de Data Mining tienen tres clases: las técnicas de modelado originado por la teoría (en que las variables se dividen en dependientes e independientes, similares a las técnicas del análisis de la dependencia o métodos explicativos del análisis multivariante), las técnicas de modelado originado por los datos (en las que todas las variables tienen inicialmente la misma importancia, similares a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante) y las técnicas auxiliares.(Pérez C. Técnicas de Análisis Multivariante de Datos Aplicaciones en SPSS)

La aplicación de todo modelo debe superar las siguientes fases:

Gráfico N° 2.1

Fases de la aplicación de modelos multivariados




Elaboración: Propia

A continuación se presenta la clasificación de las técnicas de Minerías de datos:

Tabla N° 2.2

Clasificación de Técnicas de Minería de Datos

| **Técnicas de Minería de Datos** | |
|---|---|
| **Modelo dirigido por la teoría (Técnicas Predictivas)** | Análisis de la Varianza |
| | Regresión |
| | Series Temporales |
| | Discriminante |
| **Modelo dirigido por los datos (Técnicas Descriptivas)** | Análisis Cluster |
| | Análisis Factorial |
| | Escalamiento Multidimensional |
| | Escalamiento óptimo |
| | Arboles de decisión |
| | Redes Neuronales |
| | Análisis Conjunto |
| **Técnicas Auxiliares** | Proceso Analítico de transacciones (OLAP) |
| | Reporting |

Elaboración: Propia

**Definición del Acuerdo de Basilea**

Los acuerdos de Basilea son los acuerdos de supervisión bancaria o recomendaciones sobre la legislación y regulación bancaria que son emitidos por el Comité de Supervisión Bancaria de Basilea que está compuesto por los gobernadores de los bancos centrales de las principales economías del mundo. Cada Estado o zona económica puede integrar estas recomendaciones a sus normativas. Actualmente existen tres Acuerdos de Basilea los cuales detallaremos a continuación:

**Acuerdo de Basilea I**

El acuerdo de Basilea I publicado en 1,988 estableció los principios básicos en los que debería fundamentarse la actividad bancaria como son: el capital regulatorio que debía tener una entidad bancaria en función de los riesgos que ocurren en la empresa. El acuerdo establecía que el capital mínimo de la entidad bancaria debía ser el 8% del total de los activos de riesgo (crédito, mercado y tipo de cambio sumados). El Acuerdo de Basilea I también establece el requisito de permanencia, la capacidad de absorción de pérdidas y de protección ante quiebra.

**Acuerdo de Basilea II**

El Acuerdo de Basilea II o también llamado "Nuevo Acuerdo de Capital" aprobado en el 2,004 tiene como objetivo establecer los requerimientos de capital necesarios que debe tener las entidades frente a los riesgos financieros y operativos.

Los objetivos de Basilea II son:

- Promover seguridad en el Sistema Financiero.
- Mantener un sano nivel de capital en el Sistema Financiero.
- Incrementar la competitividad bancaria.
- Constituir una aproximación más completa hacia el cálculo de riesgo.
- Plantear métodos más sensibles al riesgo.

El Nuevo Acuerdo de Capital está compuesto por tres pilares:

**Pilar I: Requerimientos Mínimos de Capital**

Propone reglas para el cálculo de los requerimientos de capital, motivando a los bancos a mejorar su administración y medición de riesgo.

Requerimiento mínimo de capital para **riesgo de crédito**. Se podrá adoptar cualquiera de los siguientes tres enfoques:

• Método Estándar (STDA).
• Método Basado en Calificaciones Internas Básico (IRBF).
• Método Basado en Calificaciones Internas Avanzado (IRBA).

Requerimiento mínimo de capital para **riesgo operacional**. Existen tres métodos para el cálculo de los requerimientos mínimos de capital:

• Método del Indicador Básico.
• Método Estándar.
• Métodos de Medición Avanzada (AMA).

Requerimiento mínimo de capital para **riesgo de mercado**. Ningún cambio desde que se incluyó en 1996 en Basilea I.

• Método estándar.
• Modelos internos.

**Pilar II – Supervisión.**

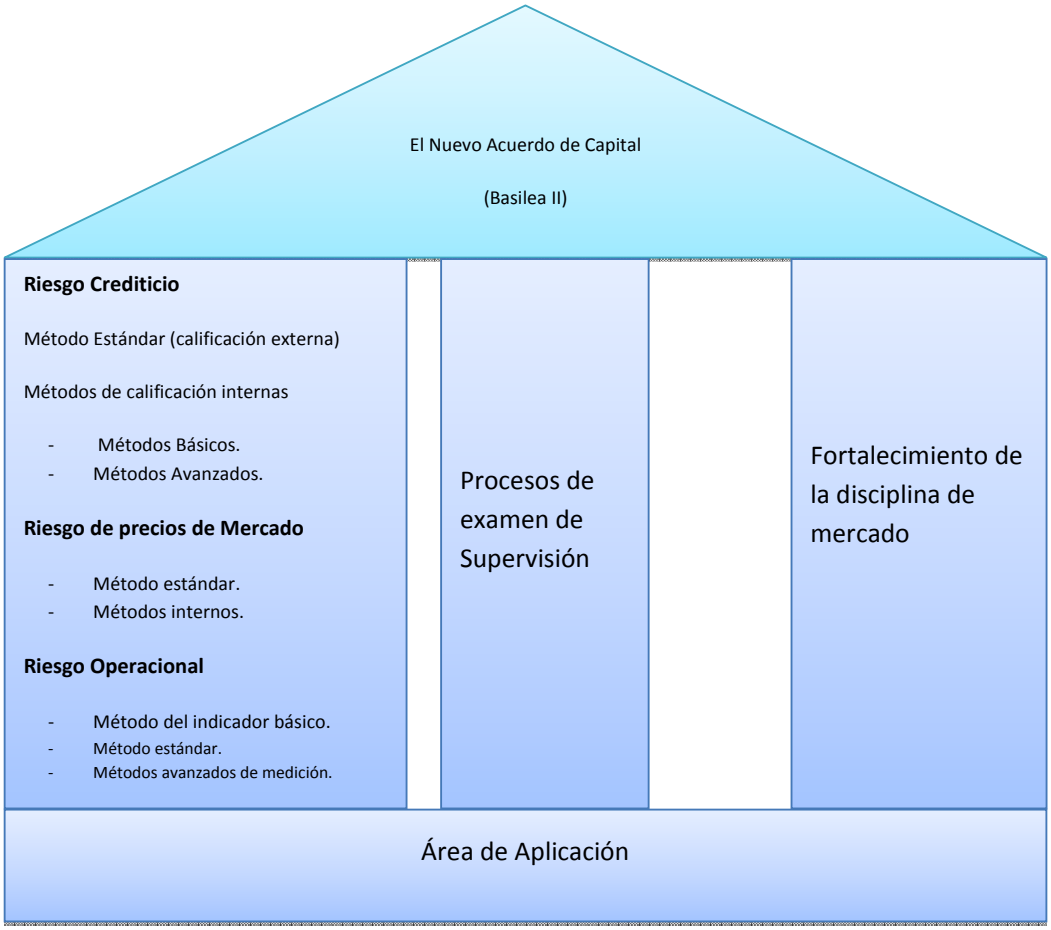
Da lineamientos para que el Supervisor promueva mejores prácticas en la administración de riesgos y se abarcan otros riesgos como el estratégico y reputacional.

**Pilar III – Disciplina de Mercado.**

Es una guía de la información que los bancos deben publicar con el fin de dar mayor transparencia a la estructura y suficiencia del capital y la exposición al riesgo de la institución.

Gráfico N° 2.2

Esquematización del Nuevo Acuerdo de Capital




Elaboración: Propia

**Diferencias entre el acuerdo de Basilea I y Acuerdo de Basilea II**

Existen ciertas diferencias entre Basilea I y Basilea II que se enumeran en la siguiente tabla:

Tabla N° 2.3

Diferencias entre el Acuerdo de Basilea I y Basilea II

| Basilea I (1,988) | Basilea II (2,003) |
|---|---|
| Estructura basada en un pilar | Se establecen 3 pilares:<br>- Requerimientos de capital<br>- Revisión de la entidad supervisora<br>- Disciplina de mercado |
| Medición del riesgo crediticio:<br>Aplicación de ponderaciones dadas por el regulador | Riesgo crediticio: Aplicación de ponderaciones externas (calificadoras) o por métodos internos. |

| | |
|---|---|
| Cálculo del Riesgo Crediticio por medio del enfoque estandarizado | Cálculo del Riesgo Crediticio mediante 3 métodos:<br>1. Estandarizado<br>2. IRB (Fundacional)<br>3. RB (avanzado) |
| Incorpora la medición del Riesgo de Mercado desde 1,996 | Permanece igual |
| No incorpora la medición del Riesgo Operativo | Incorpora la medición del Riesgo Operativo |
| Países de la OECD reciben un trato preferencial | No existe trato diferenciado para los países miembros de la OECD |
| No incluye posibilidad de requerimiento adicional por otros riesgos | El pilar 2 da la posibilidad al ente supervisor de requerir mayor capital por otros riesgos. |

Elaboración: Propia

**Acuerdo de Basilea III**

El acuerdo Basilea III se aprobó en diciembre de 2010, intentó adaptarse a la magnitud de la crisis debido a las hipotecas subprime, atendiendo a la exposición de gran parte de los bancos de todo el mundo al crecimiento excesivo de los valores presentados en los balances de los bancos y en los derivados que circulaban en el mercado. El temor al efecto dominó que pudiera causar la insolvencia de los bancos, hizo que se establecieron nuevas recomendaciones como:

- Endurecimiento de los criterios y aumento de la calidad del volumen de capital para asegurar su mayor capacidad para absorber pérdidas.
- Modificación de los criterios de cálculo de los riesgos para disminuir el nivel de exposición real.
- Constitución de colchones de capital durante los buenos tiempos que permitan hacer frente el cambio de ciclo económico.
- Introducción de un nuevo ratio de apalancamiento como medida complementaria al ratio de solvencia.

**Definición de Riesgo**

Para conocer porque se origina la necesidad de realizar los modelos de creditscoring se debe conocer primero el concepto de los distintos tipos de riesgos y para ello tomaremos las definiciones dadas por la SBS en las normativas que se establecen en el Perú:(Resolución S.B.S. N° 37 -2008: Reglamento de la Gestión Integral de Riesgos)

**Riesgo:** La condición en que existe la posibilidad de que un evento ocurra e impacte negativamente sobre los objetivos de la empresa.

**Riesgo de Crédito:** La posibilidad de pérdidas por la incapacidad o falta de voluntad de los deudores, contrapartes, o terceros obligados, para cumplir sus obligaciones contractuales registradas dentro o fuera del balance.

**Riesgo de Mercado:** Posibilidad de pérdidas en posiciones dentro y fuera de balance derivadas de fluctuaciones de los precios de mercado.

**Riesgo Operacional**: La posibilidad de pérdidas debido a procesos inadecuados, fallas del personal, de la tecnología de información, o eventos externos. Esta definición incluye el riesgo legal, pero excluye el riesgo estratégico y de reputación.

**Riesgo Estratégico**: La posibilidad de pérdidas por decisiones de alto nivel asociadas a la creación de ventajas competitivas sostenibles. Se encuentra relacionado a fallas o debilidades en el análisis del mercado, tendencias e incertidumbre del entorno, competencias claves de la empresa y en el proceso de generación e innovación de valor.

**Riesgo de Liquidez**: La posibilidad de pérdidas por incumplir con los requerimientos de financiamiento y de aplicación de fondos que surgen de los descalces de flujos de efectivo, así como por no poder cerrar rápidamente posiciones abiertas, en la cantidad suficiente y a un precio razonable.

**Riesgo de Seguro**: La posibilidad de pérdidas por las bases técnicas o actuariales empleadas en el cálculo de las primas y de las reservas técnicas de los seguros, insuficiencia de la cobertura de reaseguros, así como el aumento inesperado de los gastos y de la distribución en el tiempo de los siniestros. Se le conoce también como riesgo técnico.

**Riesgo de Reputación**: La posibilidad de pérdidas por la disminución en la confianza en la integridad de la institución que surge cuando el buen nombre de la empresa es afectado. El riesgo de reputación puede presentarse a partir de otros riesgos inherentes en las actividades de una organización.

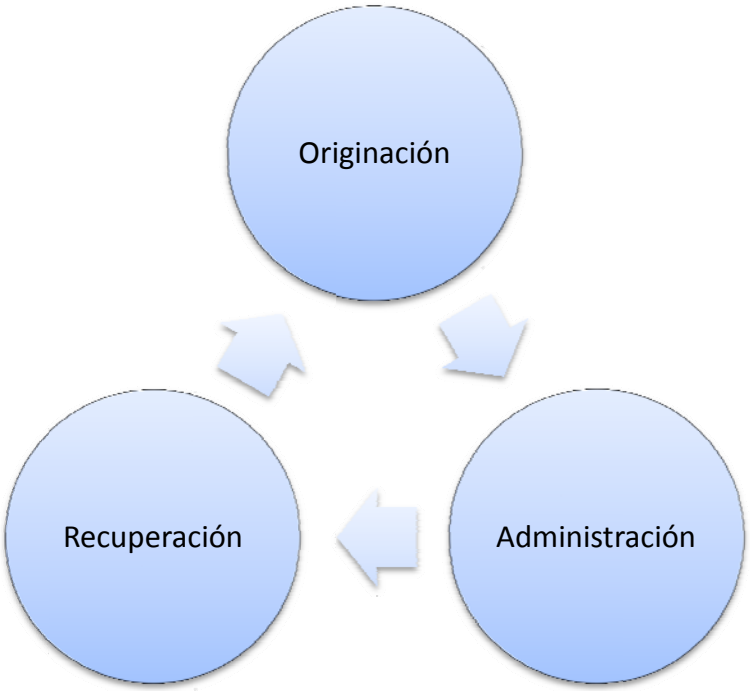
El riesgo de Crédito es el riesgo más importante en las entidades bancarias porque impacta directamente en los resultados del negocio, ya que la mala gestión de riesgo crediticio puede ocasionar perdidas a la entidad bancaria. El riesgo de crédito se administra mediante políticas de crédito más restrictivas y optimizando los procesos de otorgamiento, seguimiento y recuperación de créditos, en ello radica la importancia de tener un modelo de creditscoring que permita optimizar el proceso de otorgamiento de crédito mejorando los tiempos de otorgamiento y estandarizando los parámetros de evaluación.

**Ciclo de riesgo**

El ciclo de Riesgo está compuesto por las siguientes fases en las cuales participa el cliente y el representante comercial:(ver Nieto S. et al. 2010)

Gráfico N° 2.3

Ciclo del Riesgo




**Originación:** En esta etapa el objetivo es otorgarle un crédito por primera vez.

**Administración:** La administración consiste en detectar de manera temprana los créditos de alto riesgo y tomar las medidas necesarias para la mitigación del riesgo. Las medidas pueden ser aumentar o disminuir las líneas de créditos o las tasas de interés, reestructurar las deudas.

**Recuperación:** Consiste en tratar de recuperar a los clientes que ya no pagan sus cuotas y se le aplican medidas de cobranza a los clientes que presentan un alto puntaje en el score y a los clientes con bajo score son derivados a empresas especializadas en recaudación.

**Definición de Scoring**

**CreditScoring**

Es una herramienta que sirve para calificar o filtrar clientes de cualquier entidad que otorga crédito en base a su probabilidad de default o incumplimiento de pago (riesgo crediticio). Esta metodología crediticia determina dicha probabilidad a partir de las características personales del individuo, de su empresa y del tipo de crédito que solicita; para lo cualutiliza como información inicial el comportamiento de otros clientes que han recibido uncrédito previamente en condiciones similares.(ver Herrán C. et al 2009)

**Tipos de modelos**

**Modelo Experto**

Es el modelo que ha sido construido con información de otras instituciones y que está listo para usar. Las empresas lo adquieren a consultores externos y estos modelos son conocidos como modelos de crédito genéricos. Generalmente lo adquieren empresas que recién se inician como prestamistas debido a que no cuentan con historial de clientes.

**Modelo estadístico**

Son modelos cuya materia prima es la información de la propia empresa son conocidos como modelos *in-house*. El beneficio de realizar este tipo de modelos es que lo puedes aplicar a segmentos determinados de la cartera de la institución financiera. La otra ventaja es que se adquiere habilidad den el diseño e interpretación de los resultados además que se conserva la confidencialidad de la información.

**Ventajas y desventajas del creditscoring**

El creditscoring no reemplaza por completo el conocimiento del Analista de Crédito sin embargo tiene la suficiente capacidad de pronóstico para mejorar el proceso de evaluación crediticia.

Las principales ventajas son:

- Cuantifica el riesgo como una probabilidad.
- Los pronósticos de las probabilidades de impago son confiables.
- Es consistente debido a que dos solicitudes idénticas dan el mismo resultado.
- El proceso utilizado para el pronóstico es exacto.
- Se puede considerar un gran número de variables.
- se pueden realizar pruebas para validar el scoring.
- Revela las relaciones entre el riesgo y las características del prestatario.
- Reduce los tiempos de la gestión de cobranzas.

Las desventajas son:

- Requiere numerosos préstamos y muchos datos de cada préstamo.
- Requiere de un consultor; capaz de monitorear el sistema y hacer cambios sensibles.
- Depende de su integración con el sistema de información de la empresa.
- No tiene en cuenta variables adicionales a las que se presenta en la base de datos como por ejemplo variables de riesgos operativos.
- Puede denegar solicitudes pero no puede aprobarlas o modificarlas.
- Supone que las variables no varían.

Utilidad del creditscoring

a. Discrimina entre probables buenos y malos pagadores.
b. Asigna probabilidades de incumplimiento de pago a cada cliente.
c. Identifica las variables que afectan al riesgo crediticio, así como el efecto marginal de las mismas sobre dicha probabilidad.

d. Concentra esfuerzos de supervisión sobre los prestatarios más riesgosos.
e. Presupuesta provisiones de acuerdo al riesgo de crédito esperado.
f. Fija la tasa de interés de acuerdo al riesgo crédito esperado y se establecen tasas de interés diferenciadas según el riesgo de crédito de los clientes.
g. Define el puntaje mínimo para la aceptación de clientes buenos.
h. Establece productos de créditos personalizados y actividades de marketing dirigida a este tipo de clientes deseables.
i. Permite cumplir con las metas de la institución financiera.

Requisitos para la construcción de un Modelo de CreditScoring

1. Contar con una muestra representativa de clientes cumplidos e incumplidos.
2. Contar con una adecuada base de datos de las solicitudes de créditos.
3. Seleccionar las posibles variables explicativas del modelo de scoring en base al conocimiento experto del analista y a procedimientos estadísticos (test de significancia individual).
4. Escoger el modelo más apropiado en base a los test estadísticos sobre la bondad de ajuste o calidad predictiva del modelo.

**Tipos de Scoring**

El scoring tiene como objetivo principal generar puntajes para cada solicitud de créditos. Los tipos de scoring se determinan de acuerdo en que etapa del ciclo de riego estén:

**Score de Originación o Score de Admisión:**Este tipo de scoring se utiliza para la aceptación y rechazo de solicitudes de créditos. Las variables que utiliza son demográficas y de buró de crédito. El scoring arroja un puntaje que estima la probabilidad de incumplimiento de pago de un posible cliente con la finalidad de decidir si aceptar o rechazar al cliente con ello se consigue optimizar la tasa de aprobación de clientes buenos.

**Score de Comportamiento:** Se utiliza en la etapa de administración del ciclo de riesgo. Se encarga de predecir la probabilidad de incumplimiento de los clientes que ya son objeto de crédito en la institución. Las variables que se utilizan son las variables de comportamiento de las cuentas dentro de la propia Institución financiera. El valor agregado de este tipo de score es que permite dar seguimiento al comportamiento de los clientes lo que permitirá al área de cobranzas emplear técnicas para que un cliente siga siendo rentable para la empresa.

**Score de Cobranza:** Es el puntaje que se calcula en la parte de recuperación de cuentas para estimar la probabilidad de recuperar a un cliente. Las variables que se utilizan resultan de la combinación de variables de comportamiento y buró de crédito.

En este trabajo se consideró el score de comportamiento aunque es importante saber que las técnicas estadísticas pueden ser utilizadas de manera indistinta en los tres tipos de score.

**Score**

Es un proceso estadístico que toma información de un cliente o cuenta y la convierte en un número. El score predice la conducta de pago. Cuando estamos iniciando el desarrollo de un modelo de score uno de los primeros pasos es analizar toda la información que tengamos disponible y entender que nos está diciendo.

**El scorecard**

Una scorecard es una tabla que contiene los puntajes asignados a cada atributo de cada una de las variables usadas en el modelo estadístico. El puntaje determina la probabilidad de pago de la deuda por lo tanto a mayores puntajes corresponde una mayor probabilidad de pago. La empresa define el puntaje de separación entre clientes buenos y malos. Este puntaje de corte es llamado *cut off* es recomendado por los analistas de credit score pero será aprobado por la gerencia teniendo en cuenta las metas de la institución. (Ver Thomas et al 2002)

Existen muchas metodologías que se pueden usar para obtener una scorecard y los pasos a seguir son los siguientes:

Tabla N° 2.4

Pasos para realizar una scorecard

| | |
|---|---|
| 01 | Conformar la base de datos y agrupar los datos contenidos en la base |
| | Consiste capturar la información de los expedientes de créditos en un archivo electrónico. Adicionalmente se construye una base que contiene el comportamiento de la mora de los clientes de la base de datos anterior. Se excluyen los registros con valores extremos o los que no tienen registros, etc.Una vez que se tiene la base datos lista se procede a formar intervalos de clase o grupos de clase para cada variable (atributos de la característica). |
| 02 | Determinar los clientes buenos y malos |
| | En este paso utilizamos la base de datos que registra el comportamiento de pago de los clientes. Construimos con estos datos una matriz de transición correspondiente a la ventana de tiempo. Utilizando esta matriz determinamos los clientes buenos y malos según nuestra regla de decisión. Finalmente queda una base de datos con la variable dependiente y las variables explicativas. |
| 03 | Determinar una función de clasificación |
| | Luego de determinar a los clientes buenos y malos se procede a estimar una función para clasificar a los nuevos clientes y así poderdeterminar si se les otorga o no el crédito. La función declasificación depende de las variables que están en la solicitud que no fueron excluidas de la base de datos. Debido a la naturaleza de las variables con las que se va a calcular la función clasificadora se elige a la regresión logística para estimar a la funciónde clasificación. |
| 04 | Elaborar la scorecard |
| | Los puntajes asignados a los atributos de cada característica se calculan en función de las estimaciones de los parámetros de la función de clasificación |

| | |
|---|---|
| | obtenida con la regresión logística. Se hace una calibración de estos puntajes de acuerdo a criterios propios de la empresa. |
| 05 | Medir la eficiencia de la scorecard y definir el punto de corte |
| | En este paso se utilizan métodos estadísticos como el índice de Gini y la prueba de Kolgomorov-Smirnov para determinar la eficiencia de la clasificación del modelo de predicción.Se determina el punto de corte (cut off) que separará a las solicitudes nuevas en aceptados o rechazadas. Para ello calculamos el porcentaje de rechazo y la moratoria asociada para un score dado. |

**Determinar los clientes buenos y malos**

Se definen como clientes buenos a los clientes que pagan a tiempo sus mensualidades o permanecen en mora poco tiempo. La definición de clientes buenos y malos se hace en base a los siguientes factores:

El comportamiento de los clientes: Es la información del tiempo de pago y número de pagos vencidos del cliente.

El proceso de cobranza: Son las acciones de cobranza que se realizan para recuperar los pagos de clientes que suelen atrasarse en sus pagos.

Las metas de la institución: La institución puede definir el número de pagos vencidos para determinar a los clientes buenos y malos.

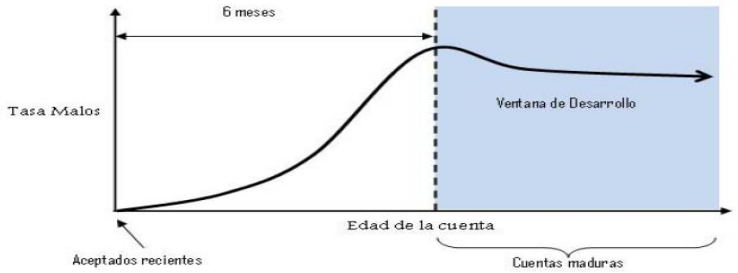
Estos factores determinan a los clientes buenos y malos. Además los clientes buenos y malos se determinan mediantes matrices de transición que son el seguimiento de clientes durante un periodo de tiempo al que se le llama ventana de muestreo.

**Ventana de muestreo**

La ventana de muestreo es un tiempo determinado después del otorgamiento del crédito en que se observa el comportamiento de los clientes. Conforme avanza el tiempo la tasa moratoria va cambiando hasta que llega el momento en que se estabiliza ese momento te indica que partir de ese tiempo ya podemos clasificar a los clientes como buenos y malos con una variación mínima y que los créditos han llegado a su madurez. En ese tiempo las cuentas indeterminadas se han reducido notablemente, debido a que ya no están en la base de datos. Es importante realizar el gráfico de la edad de la cuenta y la tasa de malos para observar el comportamiento antes mencionado como en la siguiente figura.

Gráfico N° 2.4

Ventana de Muestreo




Ventana de muestreo. Periodo de observación del comportamiento de la tasa de mora en los créditos a partir de su alta

Es importante tener cuidado al momento de tomar la muestra debido a que utilizar muestras muy antiguas no representa la realidad y las muestras muy recientes nos implica reducir el tamaño de la muestra debido a que el período de exposición se reduce y el número de cuentas para observar también. Se recomienda que se observen las cuentas de crédito un período de 12 a 18 meses anteriores a la fecha en que se realiza el estudio. Una vez que se sabe cuál es el tiempo necesario ´para llegar a la estabilización se debe definir que cuentas están contenidas en ese período y para que sean utilizadas para construir el modelo estadístico (requiere determinar el periodo de tiempo donde estarán incluidas las fechas de alta de estas cuentas).

**Obtención de la función de clasificación**

La función de clasificación depende de las variables explicativas como son: variables demográficas, las variables del comportamiento crediticio del cliente, las variables que corresponde a sus ingresos económicos, datos de su centro laboral y variables de la posición crediticia del cliente en el Sistema Financiero. Debido a la naturaleza de las variables la función clasificadora será trabajada con una regresión logística y además con un árbol de clasificación para luego comparar el poder de predicción de ambos y poder elegir el mejor.

**Metodología Estadística**

## Análisis de Regresión Logística

La predicción de variables no métricas no debe realizarse mediante una regresión múltiple, pero si desde un análisis discriminante. Sin embargo el análisis discriminante tiene algunas limitaciones importantes como son que exige los siguientes supuestos:

- Normalidad Multivariada.
- Relación lineal entre predictores.
- Homogeneidad de las varianzas – covarianzas entre grupos.

En la práctica estos supuestos no se cumplen frecuentemente por ejemplo en el caso que una de nuestras variables predictoras es categórica difícilmente se distribuirá normalmente, tampoco presentará un relación lineal con otros predictores. Tengamos en cuenta que cuando las variables predictoras no son métricas, no es recomendable emplear el análisis discriminante.

La regresión logística tiene el mismo objetivo que el análisis discriminante, predecir la pertenencia a una categoría o grupo y prácticamente de la misma manera, mediante una combinación lineal de los predictores. Pero su gran ventaja es que no plantea exigencias tan estrictas sobre las características de esos predictores, es decir:

- No asume la relación lineal entre ellos.
- No asume homogeneidad de varianzas-covarianzas.
- No asume que se distribuyan según una normal multivariada.

La semejanza entre la regresión logística y la regresión múltiple permite que la interpretación de resultados sea más fácil que la del análisis discriminante. La regresión logística además permite que las variables predictoras interactúen entre sí.

La lógica de la predicción de la regresión logística es que se evalúa la probabilidad de que una persona pertenezca a un grupo, puesto que tiene un determinado patrón de valores en las variables predictoras. La persona se asignara al grupo al que tenga mayor probabilidad de pertenecer. La variable dependiente no es el grupo en sí, sino es la probabilidad de pertenecer a un grupo dado ciertas circunstancias. Este análisis es adecuado cuando la relación entre la variable dependiente (la probabilidad) y las variables independientes (los predictores) es no lineal.

**Las ecuaciones básicas de la regresión logística:**

Cuando se trata de predecir una variable dicotómica, que adopta valores 0 ó 1 (por ejemplo, fracaso y éxito, respectivamente), su relación con los predictores es no lineal. En caso usáramos las ecuaciones de la regresión lineal correríamos el riesgo de predecir valores fuera del rango de la variable (mayores que 1 o menores que 0). Por ello, lo que se predice no es directamente la variable sino la probabilidad de que la variable adopte un cierto valor. La variable dependiente es pues una probabilidad. Para predecir una probabilidad pueden utilizarse diferentes funciones, entre las que destaca la logística. Esta función es la base del cálculo de la probabilidad $p$ que queremos predecir. Si llamamos $X_j$ a los predictores, la ecuación se escribe de la siguiente manera:

$$p = \frac{e^u}{1 + e^u} \qquad (01)$$

Donde $u$ es:

$$u = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

Esta última expresión tiene la forma de la ecuación de regresión múltiple, donde $a$ es la constante y los $b_j$son los coeficientes de los predictores $X_j$ correspondientes. **Esta expresión es conocida como logit, o logaritmo de las verosimilitudes**. La razón es la siguiente. La estimación de los parámetros implicados en el cómputo de $u$ se simplifica bastante si dividimos p por (1-p), esto es, por la probabilidad de que ocurra el otro resultado.

Teniendo en cuenta que (1-p) es

$$1 - p = 1 - \frac{e^u}{1 + e^u} = \frac{1}{1 + e^u} \qquad (02)$$

Tendremos que el cociente de probabilidades, conocido como odds ratio será

$$\frac{p}{1 - p} = e^u \qquad (03)$$

y por lo tanto tomando logaritmos tendremos

$$\ln\left(\frac{p}{1 - p}\right) = u = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p \qquad (04)$$

lo que indica que hay una relación lineal entre el cociente de probabilidades (la probabilidad de pertenecer a un grupo dividido por la probabilidad de pertenecer al otro) y los predictores.

Ahora todo el problema se reduce a estimar los parámetros. En principio pudiera pensarse que podríamos ya resolver el problema como si se tratase de estimar parámetros en regresión múltiple, sin embargo, desconocemos la probabilidad de cada sujeto de ser miembro de uno u otro grupo. Por ello, el problema reside en obtener la combinación lineal de predictores que hace máxima la probabilidad de obtener los resultados observados en la variable de agrupamiento. El procedimiento de estimación que se emplea es, por tanto, el de máxima verosimilitud. Se trata de un procedimiento de carácter iterativo en el que se comienza utilizando un conjunto de coeficientes, y se determina el ajuste de las predicciones con respecto a la variable de agrupamiento (se computan los residuos o errores de predicción). A continuación se modifican los coeficientes y se vuelven a analizar los residuos. El proceso finaliza cuando no se puede mejorar el ajuste.

## Utilidad de la regresión logística

El objetivo fundamental de la regresión logística es el mismo que el del análisis de regresión múltiple o el análisis discriminante, determinar si hay relación entre una variable predicha (usualmente la pertenencia a un grupo de sujetos, también llamada resultado) y un conjunto de predictores. La naturaleza de las variables predictoras pueden ser variables métricas o no. Si la relación entre las variables existe se debe determinar si todas los predictores son necesarios o pueden excluirse sin deteriorar la predicción de modo significativo. Así, el objetivo del análisis se concreta en hallar un modelo que incluya un conjunto de predictores más un término constante que sea significativamente superior a un modelo de referencia que habitualmente contendrá sólo la constante.

Una diferencia importante de la regresión logística con respecto a otras técnicas de clasificación como el análisis discriminante es que aquí es posible introducir términos de interacción de predictores en la ecuación.

El modelo especifica la combinación lineal de predictores que es necesario realizar para predecir el resultado. Por tanto, hay dos cuestiones importantes relativas a la contribución de cada predictor. Primera, lógicamente es necesario conocer los pesos (parámetros) de cada uno y cómo interpretarlos. La interpretación está relacionada con el cambio de resultado cuando cambian los parámetros. Segundo, cuestiones como las variables individuales afectan o no afectan a la probabilidad de obtener un resultado determinado.

En la regresión logística se utiliza para la comparación de modelos se usan medidas análogas al coeficiente de correlación múltiple o a medidas de coeficiente de determinación para medir la capacidad predictiva y la variabilidad explicada

Finalmente, en cualquier técnica de clasificación de individuos nos interesa conocer cuan eficaz es para asignar correctamente los individuos a sus categorías de pertenencia. Puesto que lo que se predice es la probabilidad de un resultado (la probabilidad de éxito), la regla de asignación es sencilla, se fija un punto de corte en la probabilidad (por ejemplo, 0,50) y se asignan a una categoría aquellos cuya probabilidad sea superior y a la otra categoría a los que estén por debajo. Contando las frecuencias de aciertos y errores de clasificación (¿cuántos errores de clasificación se producen?) podemos evaluar el ajuste global del modelo.

## Fundamentos de la regresión logística

La variable a predecir debe ser obligatoriamente categórica. Las variables predictoras pueden contener variables métricas y también categóricas para estas últimas tienen que ser codificadas de la manera apropiada. La codificación de las variables de agrupamiento debe realizarse con cuidado debido a que ello afectara a la interpretación de los coeficientes de la ecuación ya que afectara el signo de este. Respecto de los

predictores categóricos puede decirse algo semejante. Es necesario codificarlas para que puedan entrar en el análisis. La forma más común de código consiste en crear tantas variables ficticias (dummy) como categorías tenga la variable menos 1. En cada una de las variables ficticias se codifica una categoría (asignándole 1 a los sujetos que la poseen 1 y 0 a los que no), y en el conjunto quedan codificadas todas.

Es posible que no todas las variables del modelo realicen un aporte significativo al modelo de predicción una forma de evaluar la contribución de las variables consiste en evaluar sus coeficientes. Uno de los test más empleados en este sentido es el de Wald, que es un cociente entre el coeficiente y su error típico. La interpretación del cociente de Wald es sencilla, puesto que se trata de una puntuación z. El estadístico de Wald permite determinar la significación de los coeficientes del modelo.

Si el modelo produce o no un ajuste significativo en la predicción es algo que no puede conocerse considerando los coeficientes de las variables. Para evaluar el ajuste lo mejor sería realizar las predicciones y obtener una medida de la precisión alcanzada. Esta idea es la que cuantifica el logaritmo de la verosimilitud (log-likelihood), cuya expresión de cálculo es la siguiente:

$$log - likelihood = \sum_{i=1}^{N} [Y_i * \ln(p_i) + (1 - Y_i) * \ln(1 - p_i)] \quad (05)$$

Primero se computan las predicciones y a continuación se opera para obtener la suma. Sin embargo, el ajuste de un modelo debe computarse en relación al ajuste que produce otro. Cuando todas las variables han sido incluidas en la ecuación, el modelo de referencia debe ser uno que no incluya ningún predictor, solo la constante.

La comparación entre los dos modelos se realiza computando las diferencias entre el modelo que más variables incluye y el que menos. La diferencia es una chi-cuadrado que se distribuye según la diferencia entre los grados de libertad del primer y del segundo modelo. Los grados de libertad del modelo son el número de predictores incluido (teniendo en cuenta la codificación ficticia) más uno (por la constante).

$$X^2 = 2[\big(log - likelihood(F)\big) - \big(log - likelihood(C)\big)] \quad (06)$$

Consultando en las tablas de chi-cuadrado observaremos que la probabilidad de un valor como el obtenido suponiendo que las variables no aportan nada a predecir el resultado es de 0,0007, que tomaremos como valor de referencia.

**Tipos de regresión logística**

Existen varios modos de introducir las variables en la ecuación. Los predictores pueden incluirse de manera simultánea, pero también pueden introducirse de modo secuencial.

**Regresión logística simultánea**

En la regresión logística simultánea los predictores son introducidos todos a la vez. Esto implica que cada predictor es evaluado como si fuese el último en haber sido introducido en la ecuación. Recordemos que esto implica que el predictor añade a la predicción sólo lo que no comparte con los otros, lo que le diferencia de los otros. El único caso de exclusión posible es que la tolerancia de un predictor sea demasiado baja. Recordemos que la tolerancia es el complementario del coeficiente de determinación, y que una baja tolerancia se produce cuando el predictor es redundante con otro u otros. La inclusión simultánea es recomendable siempre que no se tenga claro que variables pueden ser más importantes para hacer la predicción.

**Regresión logística secuencial**

En la regresión logística secuencial se introduce o elimina un predictor en cada paso. Hay dos subtipos básicos, la regresión logística serial y la regresión logística por etapas (stepwise). La diferencia entre ambos procedimientos reside en las razones por las cuales los predictores son incluidos/excluidos de la ecuación.

**Regresión logística serial**

En la regresión jerárquica los predictores son introducidos según un criterio teórico. Esta jerarquía teórica conocida por el investigador será la que guiará, por tanto, el orden de introducción de predictores en la ecuación. Es importante caer en la cuenta de que en cada paso puede introducirse más de un predictor, pero una vez que el predictor está dentro de la ecuación no puede ser eliminado de la misma.

**Regresión logística secuencial**

En la regresión logística secuencial cada predictor es incluido o eliminado de la ecuación según cumpla los criterios estadísticos de inclusión/exclusión prefijados por el investigador. Esta aproximación es útil cuando se sospecha que hay variables más importantes que otras para la predicción, pero no se tienen hipótesis específicas sobre cuáles pueden ser o cuál puede ser el orden. Además, debe tenerse en cuenta que el problema más importante es que puede excluirse alguna variable que tenga una alta relación con la dependiente debido a su correlación con otras predictoras.

Hay varios procedimientos para tomar la decisión de introducir/excluir variables. Como sabemos ya, en los procedimientos hacia delante se comienza con la constante y se van añadiendo variables que cumplan el criterio de inclusión. En los procedimientos hacia atrás se introducen todas las variables en el primer paso y se van eliminando las que cumplan el criterio de exclusión. Hay varios métodos de selección de variables, todos basan la introducción en el estadístico de puntuación, pero difieren en lo que refiere a la exclusión, uno se basa en el cálculo condicional de los parámetros, otro en la razón de

verosimilitud y el tercero se basa en el estadístico de Wald. Los resultados pueden diferir de un método a otro.

## Limitaciones y Supuestos

### Linealidad de la función logit

La función logit es el logaritmo del cociente de probabilidades, o, lo que es lo mismo, la combinación lineal de los predictores. El supuesto implica que los predictores continuos y la función logit estén relacionados linealmente. Para ello la forma más sencilla de eliminar una interacción es mediante, una transformación logarítmica de los predictores implicados.

### Independencia de los errores

Este supuesto significa que las puntuaciones de un sujeto no son predecibles a partir de las de otro sujeto, es decir, que el valor de la variable de agrupamiento de un sujeto no puede ser predicho a partir del valor de otro sujeto. Esto ocurre cuando los sujetos han sido medidos en la variable dependiente de manera secuencial o cuando los grupos han sido igualados en variables relevantes para esto se puede considerar la variable de agrupamiento como un intrasujeto.

### Multicolinealidad

La multicolinealidad se produce cuando las variables predictoras categóricas correlacionan mucho entre sí y/o cuando las variables métricas correlacionan entre sí, lo que implica que hay predictores redundantes. La solución puede consistir en eliminar las variables que producen la multicolinealidad Para contrastarla debe examinarse la correlación (mediante análisis de frecuencias en el caso categórico) entre todos los pares de predictores.

### Número de variables y número de sujetos

Se recomienda tener un número elevado de sujetos muy superior al número de variables debido a que esto permite una buena estimación de parámetros, debido especialmente a las variables categóricas puesto que es posible que las condiciones definidas por su mezcla no contengan sujetos. De no ser así se puede ocurrir que la estimación de los parámetros puede ser muy alta y sus errores estándar también. Lo cual puede ser una consecuencia de que el algoritmo iterativo de máxima verosimilitud no encuentra una solución convergente (un conjunto de parámetros que estabilicen el error), por tanto, el número de iteraciones incrementa, y con él el valor de los parámetros.

**Puntos extremos**

La presencia de puntos extremos puede traducirse en una baja capacidad predictiva del modelo. Una solución posible es eliminarlos, pero debe tenerse en cuenta que un criterio de eliminación relajado puede incrementar artificialmente la predicción.

**El Modelo de regresión logística múltiple**

En las variables explicativas, no se establece ninguna restricción, pudiendo ser cualitativas o cuantitativas. La variable $Y$ toma los valores de 0 y 1, digamos $y = 1$ significa que es buen cliente y $y = 0$ significa que es un mal cliente. Con la regresión logística se calcula la probabilidad:

$$P(y = 1 \,|x) \tag{07}$$

es decir la probabilidad de que $y = 1$ dado los valores observados de las variables predictoras contenidas en el vector $x$.

Para justificar el modelo de regresión logística consideremos una muestra de $n$ datos donde $x_i$ es el vector de variables explicativas binarias de la forma

$$x_i^T = \left[x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}\right], i = 1,2, \dots, n \tag{08}$$

Que tiene asociada una variable de respuesta binaria dependiente $y_i$ que toma e valor de $y = 0$ si el cliente $i$ es malo, y $y = 1$ si el cliente $i$ es bueno. Sea $P(y = 1 \,|x_i) = p_i$ la probabilidad de que $y = 1$, dado el vector $x_i$ de datos observados.

La *función llamada link* (g) que relaciona $p_i$ y un modelo lineal, la función link también se conoce como la transformación logit y es el logaritmo del cociente de probabilidades de $p_i$ y $(1 - p_i)$.

$$g(p_i) = \beta_0 + \beta_1^T x_i, \tag{09}$$

Tal que

$$\beta_1^T = \left[\beta_1, \beta_2, \dots, \beta_p\right] \tag{10}$$

Vector de parámetros de coeficientes de las variables explicativas del modelo y $\beta_0$ la ordenada al origen.

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1^T x_i \tag{11}$$

El modelo en términos de $g(p_i)$ puede escribirse como $g(p_i) = \beta_0 + \beta_1^T x + \varepsilon$, con $\varepsilon$ variable aleatoria tal que $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$. La función de distribución logística dada por la transformación inversa de $g$ se escribe como

$$p_i = \frac{e^{\beta_0+\beta_1^T x_i}}{1 + e^{\beta_0+\beta_1^T x_i}} \tag{12}$$

Que satisface $0 \leq p_i \leq 1$. Así

$$1 - p_i = \frac{1}{1 + e^{\beta_0+\beta_1^T x_i}} \tag{13}$$

**Razón de probabilidades (Odds-Ratios)**

Los coeficientes del modelo logístico sirven para calcular un parámetro de cuantificación de riesgo conocido como odds ratio. El odds asociado a un evento es el cociente entre la probabilidad de que ocurra con la probabilidad de que no ocurra.

$$odds = \frac{p_i}{1-p_i} \tag{14}$$

Estos valores indican cuánto se modifican las probabilidades por unidad de cambio en las variables $x$. En efecto, de las fórmulas anteriores se deduce que

$$O_i = \frac{p_i}{1-p_i} = \exp(\beta_0) \prod_{j=1}^{k} \exp(\beta_j x_j) \tag{15}$$

Supongamos que consideramos dos elementos que tienen valores iguales en todas las variables menos en una. Sean $(x_{i1}, \dots, x_{ih}, \dots, x_{ik})$ los valores de las variables son las mismas en ambos elementos menos en la variable $h$ donde $x_{ih} = x_{jh} + 1$. Entonces el odds ratio para estas dos observaciones es:

$$\frac{O_i}{O_{i+h}} = e^{\beta_h} \tag{16}$$

e indica cuánto se modifica el ratio de probabilidades cuando la variable $x_j$ aumenta en una unidad.

**Estimación del modelo logit usando máxima verosimilitud**

Para estimar los parámetros del modelo logístico se utiliza el método de máxima verosimilitud (MV). Como $y_i$ toma dos valores, 0 con probabilidad $p_i$ y 1 con probabilidad 1-$p_i$, tiene como distribución de probabilidad de una Bernoulli.

$$P(y_i) = {p_i}^{y_i}(1-p_i)^{(1-y_i)}, y_i = 0{,}1 \tag{17}$$

La función MV para una muestra aleatoria de $n$ datos $(x_i, y_i)$ se calcula como

$$P(y_i, \dots., y_n) = \prod_{i=1}^{n} {p_i}^{y_i}(1-p_i)^{(1-y_i)}, y_i = 0{,}1 \tag{18}$$

Aplicandologaritmos

$$\log P(y) = \sum_{i=1}^{n} y_i \log(p_i) + \sum_{i=1}^{n} (1-y_i)\log(1-p_i) \tag{19}$$

Y la función log verosimilitud se escribe como

$$\log P(y) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \log(1-p_i) \tag{20}$$

Consideremos $\beta^T = (\beta_0, \beta_1, \dots, \beta_p)$ y $x_i^T = (1, x_{i1}, x_{i2}, \dots., x_{ip})$

Para escribir el modelo de la forma

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta \tag{21}$$

Ahora la ecuación 1, la sustituimos en ka ecuación 2. De aquí, obtenemos la función de verosimilitud en logaritmos en términos de los parámetros $\beta$ dada por:

$$L(\beta) = \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} \log\left(1 + e^{x_i^T \beta}\right) \tag{22}$$

Para obtener los estimadores $\beta$ de máxima verosimilitud derivamos $L(\beta)$ con respecto de cada uno de los parámetros $\beta_j$ con $j = 1{,}2, \dots, p$ e igualamos a cero.

En términos de matrices

$$\begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_j} \\ \frac{\partial \beta_1}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i(1) \\ \sum_{i=1}^{n} y_i x_{i1} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ij} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ip} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{n} (1) \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \\ \sum_{i=1}^{n} x_{i1} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \\ \vdots \\ \sum_{i=1}^{n} x_{ij} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \end{bmatrix} \tag{23}$$

cada una de estas derivadas se expresan en un vector de columna de la forma

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \left( \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right) \tag{24}$$

Igualando al vector cero

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} x_i \left( \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right) = \sum_{i=1}^{n} x_i p_i \tag{25}$$

Si $\hat{\beta}$ es el vector de parámetros que cumple el sistema, calculamos $p_i$ en términos de esos estimadores y de aquí se obtiene una estimación para $y_i$, tal que
$\widehat{y_i} = \widehat{p_i}$, así $\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} x_{ij} \widehat{y_i}$
deaquí

$$\sum_{i=1}^{n} x_{ij} e_i = \sum_{i=1}^{n} x_{ij} (y_i - \widehat{y_i}) = 0 \tag{26}$$

Donde $e_i$ representa los residuos del modelo y deben ser ortogonales al espacio de observaciones $x$, esto es similar que en la regresión estándar (mínimos cuadrados).

Observamos que el sistema de ecuaciones no es lineal en los parámetros $\beta$ y para obtener los estimadores MV es común que se utilice el método Newton-Rapson.

**Pruebas estadísticas al modelo logit**

Una de las características deseables de los modelos utilizados es que sus estimadores tengan capacidad discriminatoria. Para medir la capacidad discriminatoria se aplican diferentes técnicas de prueba que a continuación veremos.

**Deviance**

La función $D(\beta) = -2L(\beta)$ se le conoce como la desviación o deviance

$$D(\beta) = -2\sum_{i=1}^{n}\left[\log\left(1 + e^{x_i^T\beta}\right) - y_i e^{x_i^T\beta}\right] \tag{27}$$

Y en términos de probabilidades

$$D(\beta) = -2\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{28}$$

Nos dan una medida de la desviación máxima del modelo.

**Estadístico de Wald**

Para determinar si una variable debe ser incluida en un modelo porque tiene un peso significativo se aplica la prueba de estadístico de Wald. La prueba resulta de contrastar la hipótesis nula

$$H_0: \beta_i = 0$$

Contra la alternativa

$$H_1: \beta_i \neq 0$$

Con un estadístico de prueba definido como

$$\left|\widehat{w_j}\right| = \frac{\widehat{\beta_\iota}}{s(\widehat{\beta_\iota})} \tag{29}$$

Que bajo el supuesto que $H_0$ es cierto que una distribución $t$ con $n - p - 1$ grados de libertad y para muestras grandes se distribuye como una normal estándar. Se entiende que si $w_i$ es un valor alejado de cero se tendrá evidencia que $H_0$ es falsa, por lo tanto la región crítica de la prueba es dela forma $\left|\widehat{w_j}\right| > t_{\alpha/2}$, para un nivel de significación adecuado. Entendemos que si el verdadero valor del parámetro $\beta_i$ es cero la variable $x_i$ debe excluirse. Otra manera equivalente de ecribir la región crítica es usando el $p - value$ donde $p = P(t > |w_j|)$ , el $p - value$ es reportado por la mayoría de los paquetes estadísticos. La región crítica es de la forma $p < \alpha$, $\alpha$ nivel de significancia adecuado.

**Comparando modelos**

Suponga que se tiene k variables explicativas $x_1, x_2, x_3, \dots, x_k$ y se desea saber si ellas son significativas o no sin pérdida de generalidad se puede suponer que las variables a prueba son las últimas, $x_{k-s+1}, x_{k-s+2}, \dots, x_k, s < k < n$. De esta manera se está confrontando dos modelos. El primer modelo que incluye todas las variables:

$$w_1 = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-s} x_{k-s} + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k \quad (30)$$

El segundo modelo incluye solo las $k - s$ primeras variables

$$w_2 = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-s} x_{k-s} \quad (31)$$

Se realiza la prueba para contrastar la hipótesis nula $H_0$, de que las variables $x_i$ con $i = k - s + 1, \ldots, k - 1, k$ no influyen significativamente en el modelo contra la alternativa $H_1$ que dice que si influyen; esto es

$$H_0: \beta_{k-s+1} = 0 \; y \; \beta_{k-s+2} = 0 \; y \ldots . y \; \beta_k = 0$$
$$H_{10}: \beta_{k-s+1} \neq 0 \; y \; \beta_{k-s+2} \neq 0 \; y \ldots . y \; \beta_k \neq 0$$

Para encontrar la evidencia de que $H_0$ es falsa se usa la región crítica que surge del cociente de verosimilitud

$$\frac{máx \; L(H_0)}{máx \; L(H_1)} < \lambda \quad (32)$$

de esta relación se obtiene el estadístico

$$x_s^2 = 2L(H_1) - 2L(H_0) \quad (33)$$

Donde $L(H_0)$ y $L(H_1)$ son la función log-verosimilitud de cada modelo. En términos de la desvianza

$$x_s^2 = D(H_0) - D(H_1) \quad (34)$$

Si $H_0$ es cierta el estadístico sigue una distribución $x_s^2$ con $s$ grados de libertad, apra un $\alpha$ dada la región crítica es $x_s^2 > x_\alpha^2$. Esto nos da una medida de mejora entre un modelo y el otro. Nótese que cuando $s = 1$, únicamente se está probando un coeficiente del modelo de regresión y entonces se estará probando un coeficiente del modelo de regresión y entonces se estará en el mismo caso que la prueba de Wald , por lo que, la prueba de razón de verosimilitud, en este caso, es alternativa a la prueba de Wald.

**Estadístico $R^2$**

El estadístico $R^2$ sirve para estimar de manera global la influencia de todas las variables en el modelo. Un caso particular es verificar el modelo que no incluye variables de predicción y contiene únicamente $\beta_0$ contra el modelo que las incluye.

Esto es, con el cálculo de

$$R^2 = 1 - \frac{D(\hat{\beta})}{D(\widehat{\beta_0})} \quad (35)$$

Los valores extremos para $R^2$ son cero y uno. El valor $R^2 = 1$ se obtiene cuando el modelo tiene un ajuste prefecto esto es, si las observaciones $y = 1$ tienen probabilidad $p_i = 1$ y $y = 0$ probabilidad $p_i = 0$, las variables explican completamente el comportamiento de $y$. El valor de $R^2 = 0$ se obtiene cuando las variables no influyen en el modelo, no pueden predecir los valores de $y$, porque la desviación esperada para el modelo que influye todas las variables es igual a la desviación del modelo que no las incluye, o sea que $D(\hat{\beta}) = D(\widehat{\beta_0})$.

Las pruebas con estimadores de máxima verosimilitud para un modelo nos da una medida de cuan compatible es este con los datos realmente observados. Si al añadir o quitar una variable al modelo no mejora la verosimilitud o no disminuye la desviación de forma apreciable, en sentido estadístico, esta variable no se incluye en la ecuación.

**Residuos de Pearson**

Para hacer un contraste global del modelo logit podemos utilizar los residuos de Pearson. Las pruebas de hipótesis que se contrastan son:

$$H_0: El\ modelo\ es\ adecuado$$
$$H_1: El\ modelo\ no\ es\ adecuado$$

Los residuos del modelo logit están definidos como

$$e_i = \frac{y_i - \widehat{p_i}}{\sqrt{\widehat{p_i}(1-\widehat{p_i})}}, \qquad 0 < \widehat{p_i} < 1 \tag{36}$$

Si el modelo es adecuado $e_i$ tiene media cero y varianza uno, de aquí se construye el estadístico de prueba que se distribuye asintóticamente como:

$x_c^2 = \sum_{i=1}^{n} e_i^2$ que sigue una distribución $x^2$ con $n - (p+1)$ grados de libertad con $p$ variables.

**Estadistico de Hosmer-Lemeshow $C_g$**

El estadístico $C_g$ se basa en la agrupación de las probabilidades estimadas bajo el modelo de regresión $\widehat{p_1}, \ldots, \widehat{p_N}$. La idea básica es que el primer grupo estará formado

Aproximadamente por las $\frac{N}{G}$ observaciones cuyas probabilidades predichas sean más pequeñas, el segundo por los siguientes $\frac{N}{G}$ más pequeños y así sucesivamente. Los puntos de corte así generados se denominan débiles de riesgo.

La siguiente tabla muestra las frecuencias esperadas y observada, en cada uno de los grupos, utilizando el cálculo del estadístico $C_g$, denotando por $d_i, i = 1, \ldots, 10$, los débiles de riesgo de las probabilidades estimadas.

| | Respuesta | | | |
|---|---|---|---|---|
| | Y=1 | | Y=0 | |
| Grupos | Observado | Esperado | Observado | Esperado |
| $\widehat{p_J} < d_1$ | $O_{11}$ | $e_{11}$ | $O_{01}$ | $e_{01}$ |
| $d_1 < \widehat{p_J} < d_2$ | $O_{12}$ | $e_{12}$ | $O_{02}$ | $e_{02}$ |
| … | …. | … | …. | … |
| $d_9 < \widehat{p_J} < d_{10}$ | $O_{1G}$ | $e_{1G}$ | $O_{0G}$ | $e_{0G}$ |
| Total | $O_1$ | $e_1$ | $O_0$ | $e_0$ |

El número de individuos observados para los que ocurrió el suceso y para los que no ocurrió, en cada uno de los grupos es (frecuencias observadas):

$$O_{1g} = \sum_{k=1}^{ng} y_k \tag{37}$$

$$O_{0g} = \sum_{k=1}^{ng} (1 - y_k) \tag{38}$$

Siendo $n_g$ el número de observaciones en el grupo g.

Análogamente, el número esperado de individuos para los que ocurrirá el suceso y para los que no, se denotan por (frecuencias esperadas):

$$e_{1g} = \sum_{k=1}^{ng} \hat{p}(T_k) \tag{39}$$

$$e_{0g} = \sum_{k=1}^{ng} (1 - \hat{p}(T_k)) \tag{40}$$

El estadístico $C_g$ se obtiene entonces comparando estos valores observados y esperados de la siguiente forma:

$$C_g = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{(O_{kg} - e_{kg})^2}{e_{kg}} \tag{41}$$

A través de estudios de simulación se demostró que cuando J=N, si R+1<G (el número de covariables más 1 es menor que el número de grupos), bajo la hipótesis del modelo logístico, $C_g$ tiene una distribución asintótica $X^2_{G-2}$

El inconveniente en el uso de este estadístico radica en su dependencia en la elección de los puntos de corte, dando lugar a estimaciones diferentes para los mismos datos por distintos programas, pudiendo llegar incluso a darse la situación extrema de aceptación de la hipótesis nula de un ajuste adecuado por parte de algún programa y de rechazo por otro. Es por esta razón, por la que el estadístico $C_g$ se considera algo inestable, aunque la mayor parte de los softwares siguen apostando por la implementación de este test.

Regla de decisión

Si el p-value>0.05 indica que no existe diferencia entre los valores observados y estimados lo que dice que si es significativo el modelo.

Criterios para elegir el mejor modelo:

- Tenga una fuerte capacidad predictora.
- Estimación de los parámetros tenga una alta precisión.
- Modelo sea lo más sencillo posible, es decir que tenga el mínimo de variables explicativas y que satisfaga las dos condiciones anteriores.

Validación del método de clasificación

Para evaluar el método de clasificación, se utilizan datos de clientes de la misma población los cuales se conoce a que población pertenece, pero diferentes a los utilizados para estimar el modelo. Se clasifican a este conjunto de clientes mediante el modelo estimado y luego se cuentan cuántos de ellos quedan bien clasificados y cuantos quedan ml clasificados. Bajo el supuesto que el método de clasificación es adecuado, se esperaría que todos los clientes fueran bien clasificados.

En este sentido se tienen:

$O_{11} = número\ de\ clientes\ buenos\ clasificados\ como\ buenos$

$O_{12} = número\ de\ clientes\ buenos\ clasificados\ como\ malos$

$O_{21} = número\ de\ clientes\ malos\ clasificados\ como\ buenos$

$O_{21} = número\ de\ clientes\ malos\ clasificados\ como\ malos$

Tabla

Tabla de clasificación

| | Clasificados | |
|---|---|---|
| Realidad | Buenos | Malos |
| Buenos | $O_{11}$ | $O_{12}$ |
| Malos | $O_{21}$ | $O_{22}$ |

Mientras que se considera como valores esperados:

$E_{11} = número\ esperado\ de\ clientes\ buenos\ clasificados\ como\ buenos\ y$
$es\ igual\ al\ número\ de\ clientes\ buenos\ en\ la\ muestra\ de\ validación$
$E_{12} = número\ esperado\ de\ clientes\ buenos\ clasificados\ como\ malos$
$E_{21} = número\ esperado\ de\ clientes\ malos\ clasificados\ como\ buenos$
$E_{21} = número\ esperado\ de\ clientes\ malos\ clasificados\ como\ malos$

Usando la estadística:

$$x_c^2 = \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{42}$$

Se puede validar el método de clasificación. Mientras más grande es el valor de esta estadística, es menor la capacidad clasificatoria del modelo.

**Construcción de la scorecard**

Los puntajes del score son el resultado de un re-escalamiento y una traslación del modelo logístico, dado por la ecuación

$$\begin{aligned} Score &= Offset + Factor * \ln(odds) \\ &= Offset + Factor * (\widehat{\beta_0} + \sum \hat{\beta} woe_{ij}) \end{aligned} \tag{43}$$

donde Offset es un término de traslación (o compensación) y Factor es un término de re-escalamiento. Offset y Factor deben satisfacer condiciones impuestas por la empresa de crédito. Este procedimiento permite la estandarización del score para que diferentes scorecard puedan ser comparadas. Se acostumbra a calibrar la scorecard de tal manera que cada cierto incremento de puntaje $P_0$, se obtenga el doble de la relación good/bad. Para ello se resuelve el sistema de ecuaciones

$$\begin{aligned} Score &= Offset + Factor * \ln(odds) \\ Score + P_0 &= Offset + Factor * \ln(2 * odds) \end{aligned} \tag{44}$$

Cuya solución es:

$$Factor = \frac{P_0}{\ln(2)}\ y\ Offset = Score - Factor * \ln(odds) \tag{45}$$

Determinación del punto de corte o Cut off

Cuando se tiene los datos de un nuevo solicitante, se calcula su score y con el resultado se decide si se le otorga o no el crédito. Se elige un punto "a" llamado punto de corte o Cut Off tal que si Score>a se otorga crédito, en caso contrario si $Score \leq a$ se rechaza la solicitud, es importante determinar el valor que optimiza la decisión. En esta sección presentamos dos maneras de estimar el punto de corte. La primera forma es suponer que si la probabilidad de ser buen cliente está por arriba de un medio, se aprueba el crédito y si está por debajo, se rechaza; esto es, para aprobar un crédito se utiliza la desigualdad.

$$\frac{e^{\beta_0+\beta_1^T x}}{1+e^{\beta_0+\beta_1^T x}} = \hat{p} > \frac{1}{2} \tag{46}$$

De aquí se sigue que:

$$e^{\beta_0+\beta_1^T x} > \frac{1}{2}\left(1+e^{\beta_0+\beta_1^T x}\right) \rightarrow \frac{1}{2}e^{\beta_0+\beta_1^T x} > \frac{1}{2} \rightarrow e^{\beta_0+\beta_1^T x} > 1 \tag{47}$$

$$\rightarrow \widehat{\beta_0} + \widehat{\beta_1^T} x > 0 \rightarrow a = Offset \tag{48}$$

La segunda forma de obtener el punto de corte es calcular el score para todas las cuentas de la base, estos puntajes se ordenan y el valor de a satisface la ecuación:

$$\frac{\#\{Score|Score < a\}}{\#\{Score\}} = una\ proporción \tag{49}$$

Una proporción con el valor seleccionado por la empresa.

Prueba de diferencias de dos poblaciones

Una vez que se ha calculado el score con la fórmula estimada por la regresión logística se quiere determinar si estos valores calculados en la muestra identifican bien a que grupo pertenecen. Mientras mayor sea la diferencia de los puntajes score de los grupos mayor será la capacidad discriminante del modelo usado. Entre las técnicas para determinar la diferencia entre los puntajes del score en los grupos de buenos y malos clientes están: el índice de Gini, la divergencia y la prueba de Kolgomorov-Smirnov.

Índice de Gini

El índice de Gini es uno de los más utilizados para medir la desigualdad entre dos poblaciones. En este caso se utiliza para medir la desigualdad entre buenos y malos. Teóricamente la curva de Lorenz de las funciones de distribución $F(x) y\ G(x)$ es el subconjunto del producto cartesiano dado por

$$L(F,G) = \{(u,v) | u = F(x) y\ v = G(x); con\ x\ \in\ \mathbb{R}\}$$

Definimos a $F\ y\ G$ como las funciones de distribución teóricas asociadas a los clientes malos y buenos respectivamente, donde $x$ es el puntaje score. Si el puntaje de score para buenos es mayor que el puntaje score para malos, la curva de Lorenz de $F\ y\ G$ es cóncava hacia arriba. Se ve que si $F(x) = G(x)$ entonces $L(F,G)$ describe la recta u=v con $u\ \in (0;1)$ entre las distribuciones $F\ y\ G$. El índice de Gini resulta de la razón entre el Área A y el área del triángulo delimitado por la identidad, el eje horizontal $u$ y la recta $u = 1$.

Índice de Gini con observaciones agrupadas

Cuando se desconocen las funciones de distribución $F(x)\ y\ G(x)$, pero se cuenta con una muestra aleatoria de cada una de estas dos distribuciones empíricas de tamaño $n_1\ y\ n_2$ respectivamente se puede estimar la curva de Lorenz y por lo tanto el índice de Gini. Para hacer esto primero se define una partición de $\mathbb{R}$ dada por $x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_k$, luego se obtiene los estimadores $F\ y\ G$ en los puntos $x_i$ de la siguiente manera:

$$\hat{F}(x_i) = \frac{\#\ de\ elementos\ en\ la\ muestra\ 1 \leq x_i}{n_1} \tag{50}$$

$$\hat{G}(x_i) = \frac{\#\ de\ elementos\ en\ la\ muestra\ 2 \leq x_i}{n_2} \tag{51}$$

La estimación de la curva de Lorenz de $F(x)\ y\ G(x)$ es igual a la unión de los segmentos de recta que unen los puntos $(\hat{F}(x_{i-1})\ , \hat{G}(x_{i-1}))$ y $(\hat{F}(x_i)\ , \hat{G}(x_i))$ es igual.
El área debajo de la curva de Lorenz estimada para un intervalo tiene la forma de un trapecio y la calculamos como:

$$A_i = \frac{\left(\hat{F}_i - \widehat{F_{i-1}}\right)\left(\hat{G}_i - \widehat{G_{i-1}}\right)}{2} \tag{52}$$

El área total por debajo de la curva de Lorenz estimada es $(\hat{F}(x_i)\ , \hat{G}(x_i))$

$$A = \sum_{i=2}^{k} A_i \tag{53}$$

El índice de Gini estimado se calcula como

$$Gini = \frac{\frac{1}{2} - A}{\frac{1}{2}} \tag{54}$$

Regla de decisión

| Valor Gini | Calidad de Clasificación del modelo |
|---|---|
| <0.35 | Bajo |
| 0.35 – 0.55 | Regular |
| 0.55-0.70 | Bueno |
| >0.70 | Muy Bueno |

Divergencia

La divergencia mide la diferencia entre las medias de dos distribuciones estandarizadas usando las varianzas y tiene la siguiente expresión:

$$Divergencia = \frac{2(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{55}$$

Cuando construimos un modelo logístico que clasifica dos poblaciones, se espera que los dos grupos estén estadísticamente bien separados; esto es, la diferencia entre sus medias sea importante. Entre más pequeña la divergencia nos estará diciendo que la distribución de cada población es parecida y no sabremos diferenciar un grupo del otro, es decir para un mismo puntaje e score tendremos cantidades similares de buenos y malos. La divergencia debe ser mayor de 0.95 para tener poblaciones estadísticamente separadas.

Test de Kolgomorov-Smirnov

Una de las pruebas no paramétricas para la bondad de ajuste es el test de Kolgomorv-Smirnov. Si se desea probar que dos muestras independientes proviene de la misma distribución utilizamos la prueba de Kolgomorov-Smirnov también conocida como la prueba de K-S. El estadístico de prueba se calcula como la máxima diferencia absoluta entre sus distribuciones empíricas, entonces se busca detectar las discrepancias existentes entre las frecuencias relativas acumuladas de las dos muestras de estudio. Estas diferencias están determinadas no solo por las medias sino también por la

dispersión, simetría o la oblicuidad. La prueba se construye sobre la hipótesis nula y la hipótesis alternativa:

$$H_0: las\ distribuciones\ poblacionales\ son\ iguales$$
$$H_1: las\ distribuciones\ poblacionales\ son\ diferentes$$

Para esta prueba se requiere tener dos muestras de una variable aleatoria continua, o al menos de escala ordinal. Con los datos agrupados en k categorías o intervalos se calculan las frecuencias relativas acumuladas $\widehat{F_i}\ y\ \widehat{G_i}$ con $i = 1,2, \dots, k$ que corresponden a las dos muestras de tamaño $n_1$ y $n_2$ respectivamente.

Calculamos entonces las diferencias entre las frecuencias relativas acumuladas. El estadístico esta dado como la máxima diferencia de las distribuciones de frecuencias relativas acumuladas.

$$D_{máx} = máx_{1 \le i \le k} \left| \widehat{F_i} - \widehat{G_i} \right| \tag{56}$$

Se selecciona aquel intervalo de clase que tenga mayor desviación absoluta D. El valor crítico es calculado como

$$D_{crítico} = 100K \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{57}$$

donde$n_1\ y\ n_2$ son los tamaños de las muestras y $K$ es el valor obtenido de tabla de Kolgomorov-Smirnov con $n_1 + n_2 - 2$ grados de libertad a un nivel de significancia dado. Si la desviación observada es menor que la desviación crítica tabulada se acepta $H_0$, es decir que los datos observados no presentan diferencia significativa entre las poblaciones. La función de distribución no discrimina las poblaciones, es la misma para ambas. Se rechaza $H_0$ si $D_{máx} > D_{crítico}$, la distribución no es la misma para cada población, la prueba marca que hay discriminación entre las dos poblaciones.

Regla de decisión

| Valor K-S | Calidad de Clasificación |
|---|---|
| <0.20 | Malo |
| 0.20-0.40 | Aceptable |
| 0.40-0.60 | Bueno |
| 0.60-0.75 | Muy Bueno |
| >0.75 | Sospechoso |

Análisis de Residuos

El análisis de residuos en una regresión logística no es tan sencillo como en una regresión lineal ya que los valores que toma la respuesta son 1 o 0. Los residuos ordinarios no se distribuirán normalmente y su distribución bajo la suposición de que el modelo ajustado es correcto no se conoce. El gráfico de residuos ordinarios versus los valores ajustados no serán informativos. En la regresión lineal un supuesto clave es que la varianza del error no depende de la media condicional E (Y | X). Sin embargo, en la regresión logística, hay errores binomiales y, como resultado, la varianza del error es una función de la media condicional como V (Y | X) = θ (1-θ). Por lo tanto, el residuo ordinario puede hacerse más comparable dividiéndolos por el error estándar estimado de Yi que se conoce como residuo de Pearson denotado por pri y definido como:

$$pr_i = \frac{\hat{\varepsilon}}{\sqrt{\hat{\theta}_i\left(1-\hat{\theta}_i\right)}} = \frac{\left(Y_i-\hat{\theta}_i\right)}{\sqrt{\hat{\theta}_i\left(1-\hat{\theta}_i\right)}} \qquad (58)$$

Los residuos de Pearson están directamente relacionados con la estadística de bondad de ajuste de chi-cuadrado de Pearson. El cuadrado de Pearson residual mide la contribución de cada respuesta binaria a la estadística de prueba de Chi-cuadrado de Pearson, pero el estadístico de prueba no sigue una distribución aproximada de chi-cuadrado para datos binarios sin replicados. Los residuos de Pearson no tienen una variación unitaria, ya que no se ha tenido en cuenta la variación inherente en el valor ajustado. Un mejor procedimiento es normalizar aún más los residuos ordinarios por su desviación estándar estimada que se denomina residuos de Pearson estudentizados. La desviación estándar es aproximada por:

$$\sqrt{\hat{\theta}_i\left(1-\hat{\theta}_i\right)\left(1-h_{ii}\right)} \qquad (59)$$

Los residuos estudentizadosde Pearson son principalmente útiles en la identificación de observaciones influyentes y en la información sobre la influencia de un caso, mientras que los residuos de Pearson no lo hacen. Los casos más influyentes, con altos resultados, resultan en altos residuosestudentizados de Pearson. Los residuos estudentizadosde Pearson siguen aproximadamente la distribución normal estándar para muestras grandes (n≥30) y pueden usarse como una distribución aproximada de chi-cuadrado.

Los residuos desviación es otro tipo de residuo. Mide el desacuerdo entre cualquier componente de la probabilidad log del modelo ajustado y el componente correspondiente de la probabilidad log que resultaría si cada punto estuviera ajustado exactamente.Los residuos de desviación también pueden ser útiles para identificar

posibles valores atípicos o mal especificados en el modelo. El residuo de desviación para el i-ésimo caso se define como la raíz cuadrada firmada de la contribución de ese caso a la suma de la desviación del modelo como:

$$dr_i = \text{sign}\left(Y_i - \hat{\theta}_i\right)\left\{-2\left[Y_i \ln\left(\hat{\theta}_i\right) + (1 - Y_i)\ln\left(1 - \hat{\theta}_i\right)\right]\right\}^{1/2} \tag{60}$$

Una buena manera de ver el impacto de varios residuos es graficarlos contra las probabilidades predichas o simplemente con los números de casos.

La Distancia de Cook mide la influencia en la estimación de β. Aquellas observaciones evaluadas en la distancia de Cook y cuyovalor supere a 1 se consideran influyentes.

$$\Delta\hat{\beta}_i = \left(\hat{\beta} - \hat{\beta}_{(-i)}\right)' \left(X'\hat{W}X\right)\left(\hat{\beta} - \hat{\beta}_{(-i)}\right) = \frac{pr_i^2 h_{ii}}{(1 - h_{ii})^2} = \frac{spr_i^2 h_{ii}}{(1 - h_{ii})} \tag{61}$$

El gráfico de $\Delta\widehat{\beta}_i$contra $\hat{p}_i$ es de gran utilidad para detectar lasposibles observaciones influyentes.

## 2.3 Formulación de Hipótesis

### 2.3.1 Hipótesis Principal

El modelo de regresión logística permita clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria, según su probabilidad de impago, con una proporción mayor a 50%.

### 2.3.2 Hipótesis Secundaria

**Hipotesis**

Las variables tipo de docente, régimen de pensión, edad, estado civil, ingresos económicos y tipo de servidor son las más significativas en el modelo de regresión logística para clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria, según su probabilidad de impago.

## 2.4 Variables y Operacionalización

| N° | Variables | Tipo de Variable | Escala de medición | Categorías |
|---|---|---|---|---|
| 01 | Edad | Cuantitativa | Ordinal | Numérico |
| 02 | Género | Cualitativa | Nominal | 1. Femenino<br>2. Masculino |
| 03 | Nivel Ingresos | Cuantitativa | Ordinal | Numérico |
| 04 | Tenencia de negocio | Cualitativa | Nominal | S. SI<br>N. No |
| 05 | Tenencia de vivienda propia | Cualitativa | Nominal | 1. Alquilada<br>2. Familiar<br>3. Propia<br>4. Otro |
| 06 | Tipo de docente | Cualitativa | Nominal | DO. Docente<br>CE. Cesantes |
| 07 | Estado Civil | Cualitativa | Nominal | 1. Soltero<br>2. Casado<br>3. Divorciado<br>4. Viudo |
| 08 | Régimen de Pensión | Cualitativa | Nominal | AFP<br>DL. 19990<br>D.L. 20530<br>Otros |
| 09 | Fecha de Nombramiento | Cualitativa | Ordinal | Fecha |
| 10 | Fecha de Cese | Cualitativa | Ordinal | Fecha |
| 11 | Refinanciado | Cualitativa | Nominal | S. SI |

| | | | | |
|---|---|---|---|---|
| | | | | N. NO |
| 12 | Fondo Solidario Contingencia | Cualitativa | Nominal | S. SI<br>N. NO |
| 13 | Monto de la Cuenta Individual | Cuantitativa | Ordinal | Numérica |
| 14 | Tipo de Beneficio Previsional Entregado | Cualitativa | Nominal | 1. Fallecimiento<br>2. Invalidez<br>3. Cese |
| 15 | Calificación Interna | Cualitativa | Ordinal | 0. Normal<br>1. CPP<br>2. Deficiente<br>3. Dudosa<br>4. Perdida |
| 16 | Calificación Externa SBS | Cualitativa | Ordinal | 0. Normal<br>1. CPP<br>2. Deficiente<br>3. Dudosa<br>4. Perdida |
| 17 | Número de Entidades SBS | Cuantitativa | Ordinal | Numérica |
| 18 | Saldo de Deuda Externa SBS | Cuantitativa | Ordinal | Numérica |
| 19 | Monto Otorgado | Cuantitativa | Ordinal | Numérica |
| 20 | Fecha ultimo Crédito Otorgado | Cualitativa | Ordinal | Fecha |
| 21 | Tipo Crédito del último Crédito | Cualitativa | Nominal | 1. Consumo<br>2. Vivienda<br>3. Express |
| 22 | Número de cuotas vencidas | Cuantitativa | Ordinal | Numérica |
| 23 | Número de cuotas diferidas vencidas | Cuantitativa | Ordinal | Numérica |
| 24 | Saldo Interno total a la fecha | Cuantitativa | Ordinal | Numérica |
| 25 | Número de Créditos en la Institución | Cuantitativa | Ordinal | Numérica |

Nuestra variable respuesta será

| N° | Variables | Tipo de variable | Escala de medición | Categorías |
|---|---|---|---|---|
| 01 | Variable Respuesta | Cualitativa | Nominal | 0. Malo<br>1. Bueno |

## 2.5 Marco Legal

En la "Ley Nº 30114: Ley de Presupuesto del Sector Público para el Año Fiscal 2014" se autoriza en la Cuadragésima Primera Disposición Complementaria Final lo siguiente:

> ***CUADRAGÉSIMA PRIMERA***. *Autorizase a las entidades del sector público, a partir de la vigencia de lapresente disposición, a afectar la planilla única de pago con conceptos expresamente solicitados y autorizadospor el servidor o cesante, vinculados, únicamente, a operaciones efectuadas por fondos y conceptos debienestar y por entidades supervisadas y/o reguladas por la Superintendencia de Banca, Seguros y AFP, conexcepción de los sujetos obligados a reportar a la Unidad de Inteligencia Financiera (UIF) exclusivamente para fines de lavado de activos y financiamiento del terrorismo con arreglo a la Ley 29038, los que se aplican luego de otrosdescuentos de ley y mandato judicial expreso, de ser el caso, debiendo contar con la conformidad de las oficinasgenerales de administración o las que hagan sus veces en las entidades públicas.*
>
> *Para tal fin, se tiene en cuenta que el servidor o cesante reciba, por lo menos, el cincuenta por ciento (50%) de su remuneración, compensación económica o pensión neta mensual, según corresponda. Este porcentaje puede ser reajustado mediante decreto supremo refrendado por el Presidente del Consejo de Ministros y el ministro deEconomía y Finanzas, a propuesta de la Autoridad Nacional del Servicio Civil, con opinión de la Superintendencia de Banca, Seguros y AFP. Bajo la misma formalidad, en un plazo que no exceda de treinta días hábiles de la entradaen vigencia de la presente disposición, se aprueban lasnormas para que las entidades públicas adecúen los descuentos que realiza actualmente en la planilla únicade pago al citado porcentaje, así como los criterios,plazos, modalidades permitidas para los descuentos yotras necesarias para su aplicación.*

En el "Decreto Supremo N° 010-2014: Aprueba normas reglamentarias para que las entidades públicas realicen afectaciones en la Planilla Única de Pagos" se establecen las siguientes disposiciones:

> ***TÍTULO II***
> ***DE LOS DESCUENTOS***
>
> ***Artículo 3.- Determinación de base de cálculo***
>
> *Para efecto de lo establecido en la Cuadragésima Primera Disposición Complementaria Final de la Ley N° 30114, Ley de Presupuesto del Sector Público para el Año Fiscal 2014, entiéndase que el porcentaje al que se refiere el segundo párrafo de dicha Disposición se calculará sobre la base del monto neto total que mensual y permanentemente perciba el servidor o cesante y que constituya ingreso de su libre disponibilidad, independientemente de la naturaleza de los conceptos que dicho monto pudiera comprender.*
>
> *Se entiende por monto neto el que resulta luego de descontar, de la remuneración, compensación económica o pensión, los montos derivados de mandatos judiciales o legales expresos, incluyendo dentro de estos últimos los que pudieran corresponder por concepto de cuotas y descuentos sindicales.*

*De conformidad con lo establecido en el primer párrafo de este artículo, las sumas que el servidor perciba del CAFAE, por concepto de incentivo único, se entienden incluidas dentro de la referida base de cálculo.*

*La afectación de la planilla para la atención de las solicitudes formuladas por los servidores o cesantes al amparo de lo establecido en la referida Cuadragésima Primera Disposición Complementaria Final de la Ley N° 30114 no podrá afectar, en ningún caso, los montos que pudieran estos percibir de manera ocasional o eventual, tales como aguinaldos, gratificaciones o conceptos de naturaleza similar. Dichos montos no se considerarán para la determinación de la base de cálculo a la que este artículo se refiere.*

***Artículo 4.- Alcances de la solicitud***

*El servidor o cesante podrá solicitar la afectación de la planilla única de pagos sólo para efectuar, a través de ella, el pago de obligaciones asumidas por dicho servidor o cesante con aquellos Fondos de bienestar y entidades supervisadas y/o reguladas por la Superintendencia de Banca, Seguros y AFP incluidos en los registros a los que se refiere la Primera Disposición Complementaria y Final de estas normas reglamentarias.*

*En el caso de los Fondos de bienestar, la afectación procederá únicamente para la atención de obligaciones del servidor o cesante vinculados a los conceptos de bienestar siguientes: alimentación, salud, vivienda, educación, sepelio o esparcimiento.*

***Artículo 5.- Prelación***

*Al momento de efectuar la afectación de la planilla única de pagos solicitada por el servidor o cesante, la entidad considerará, en primer término, aquella que tuviera relación con la atención de las obligaciones asumidas por estos frente a los Fondos de bienestar y, sólo después, podrá considerar las relacionadas con créditos otorgados por las entidades supervisadas y/o reguladas por la Superintendencia de Banca, Seguros y AFP.*

***Artículo 6.- De las afectaciones a la planilla única***

*Las afectaciones a la planilla única de pago para la atención de obligaciones contraídas por el servidor o cesante con las entidades supervisadas y/o reguladas por la Superintendencia de Banca, Seguros y AFP, y/o con los Fondos de Bienestar, se les aplicará, según corresponda, las siguientes disposiciones:*

*6.1.- La afectación a la planilla única de pago para la atención de obligaciones contraídas por el servidor o cesante con las entidades supervisadas y/o reguladas por la Superintendencia de Banca, Seguros y AFP a las que se refiere el artículo 4 de este Decreto Supremo, podrá ser efectuada por las entidades del Sector Público sólo cuando éstos hubieran solicitado y autorizado dicha afectación para amortizar, a través de ella, cuotas de créditos de consumo no revolventes contraídos bajo un esquema de cuota fija y por un plazo máximo de amortización total de setenta y dos (72) meses.*

*Se entiende como créditos de consumo no revolventes aquellos a los que se refiere el numeral 2 del Capítulo I del Reglamento para la Evaluación y Clasificación del Deudor y la Exigencia de Provisiones aprobado mediante la Resolución SBS N° 11356-2008 y sus normas modificatorias.*

*6.2.- La afectación a la planilla se hará considerando que el descuento a efectuar al servidor o cesante por este concepto sea por un monto tal que no impida que éste reciba cuando menos el cincuenta por ciento (50%) del monto*

*neto que mensual y permanentemente le correspondería percibir de acuerdo a lo establecido en el artículo 3.*

*Si durante la vigencia del período de afectación autorizado por el servidor o cesante dicho monto neto se viera reducido por efecto de un mayor descuento por mandato legal o judicial, el porcentaje señalado en el párrafo precedente podrá reducirse al cuarenta por ciento (40%) de dicho monto.*

*6.3.- Con la finalidad de garantizar el cumplimiento de la condición establecida en los párrafos precedentes, la entidad afectará la planilla en un monto menor al autorizado por el servidor o cesante.*

*En dicho supuesto, de existir concurrencia de créditos de consumo no revolventes, la entidad efectuará la afectación para atender primero a aquel al que se refiera la más antigua de las solicitudes presentadas por el servidor o cesante. Tratándose de solicitudes de la misma antigüedad, la afectación de la planilla se efectuará proporcionalmente.*

*6.4.- La entidad supervisada y/o regulada por la Superintendencia de Banca, Seguros y AFP a que se refiere la Cuadragésima Primera Disposición Complementaria de la Ley N° 30114 podrá ampliar el plazo de pago de los créditos, de acuerdo con lo pactado con el servidor o cesante y de conformidad con la disposiciones emitidas por dicha Superintendencia, siempre que dicha ampliación no incremente la carga financiera mensual del servidor o cesante. En este supuesto, para el descuento por planilla el crédito mantendrá la prelación que tenía antes del cambio de las condiciones contractuales.*

***Artículo 7.- De las afectaciones vigentes***

*La afectación de la planilla única de pago de las entidades del Sector Público por la amortización de los créditos desembolsados antes de la entrada en vigencia de la presente norma, seguirá efectuándose en los montos y condiciones solicitados y autorizados por el servidor o cesante. La entidad supervisada y/o regulada por la Superintendencia de Banca, Seguros y AFP a que se refiere la Cuadragésima Primera Disposición Complementaria de la Ley N° 30114 podrá ampliar el plazo de pago de los créditos antes mencionados, de acuerdo con lo pactado con el servidor o cesante y de conformidad con la disposiciones emitidas por dicha Superintendencia, siempre que dicha ampliación no incremente la carga financiera mensual del servidor o cesante. En este supuesto, el crédito mantendrá la prelación que tenía antes de la entrada en vigencia de la Cuadragésima Primera Disposición Complementaria de la Ley N° 30114.*

*Cualquier nuevo desembolso que las entidades supervisadas y/o reguladas por la Superintendencia de Banca, Seguros y AFP efectuaran con cargo a las líneas asociadas a un crédito de consumo otorgado antes de la vigencia de la presente norma, y que afecte la planilla única de pago de las entidades del Sector Público, será considerado como un nuevo crédito y le serán por tanto aplicables todas las condiciones señaladas en el artículo anterior.*

## 2.6 Glosario de términos

En la Resolución SBS N° 11356-2008 (2,008) establece las siguientes definiciones acerca de los tipos de créditos:

**Créditos directos**

Representa los financiamientos que, bajo cualquier modalidad, las empresas del sistema financiero otorguen a sus clientes, originando a cargo de éstos la obligación de entregar una suma de dinero determinada, en uno o varios actos, comprendiendo inclusive las obligaciones derivadas de refinanciaciones y reestructuraciones de créditos o deudas existentes.

**Créditos de Consumo No-Revolvente**

Son aquellos créditos no revolventes otorgados a personas naturales, con la finalidad de atender el pago de bienes, servicios o gastos no relacionados con la actividad empresarial.

Son aquellos créditos reembolsables por cuotas, siempre que los montos pagados no puedan ser reutilizables por el deudor. En este tipo de crédito no se permite que los saldos pendientes fluctúen en función de las propias decisiones del deudor.

**Créditos de Convenio de Descuento por Planilla**

Son aquellos créditos orientados a personas naturales quecuenten con ingresos de tipo dependiente como boletas de pago yque la empresa en la que trabajan se compromete, medianteconvenio suscrito, a hacer efectivo los descuentos para girarlos a favor de la empresa financiera.

**Superintendencia**

Superintendencia de Banca, Seguros y Administradoras Privadas de Fondos de Pensiones.

**Incumplimiento de Pago**

El incumplimiento de pago debe definirse con cautela, por lo que es necesario identificar todo atraso que conlleve un coste para la organización. Para ello se han de verificar las siguientes condiciones:

- El atraso percibido ha de ser real y no estimado, según fechas concretas marcadas en la contratación de crédito, en función del método estipulado para su amortización por las partes contratantes.
- El atraso ha de producirse en, al menos, una cuota de amortización del microcrédito.

# CAPÍTULO III: MARCO METODOLÓGICO

## 3.1 Tipo, Nivel y Diseño de la investigación

La presente investigación es de tipo aplicado, cuantitativa, descriptiva, predictiva y de corte transversal debido a que permitirá estimar el modelo scoring el cual pronostique la probabilidad de incumplimiento de impago.

## 3.2 Población y Muestra

La población está compuesta por la base de clientes con deuda en la Entidad Financiera en el departamento de Lima durante el año 2013 y 2014, se eligió el departamento de Lima por ser el que presentan mayor cantidad de clientes. El tamaño de la población en el estudio asciende a 60,234 clientes. Los casos admitidos han de contener toda la información de las variables explicativas en caso contrario se eliminan del análisis.

Debido a que la población se encuentra dividida en dos clases clientes malos 11.90% y clientes bueno 88.1% se utilizara el muestreo undersampling que consiste en seleccionar todas las observaciones Target y un número de observaciones no taget. En este caso al solo contar con 7,167 clientes malos se debe elegir de manera aleatoria un número igual de clientes buenos esto también se conoce como submuestreo.

Tabla N° 3.1

DISTRIBUCIÓN DE FRECUENCIA DE LA PROBABILIDAD DE INCUMPLIMIENTO DE PAGO DEL CLIENTE

| Categoría | Población de Estudio | |
|---|---|---|
| | Clientes | Porcentaje |
| Mal Pagador | 7,167 | 11.90% |
| Buen Pagador | 53,067 | 88.10% |
| Total | 60,234 | 100.00% |

Fuente: Base de datos de la población de estudio

Elaboración: Propia

# CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

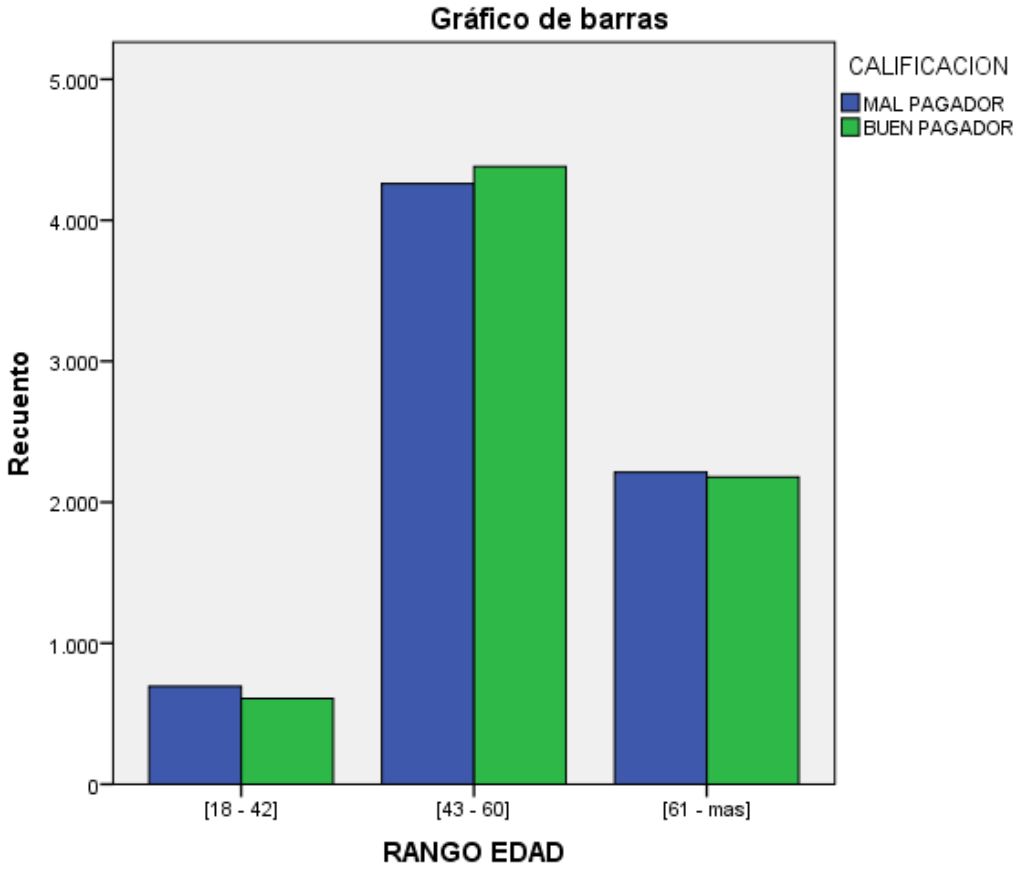
## 4.1 Resultados Preliminares

### 4.1.1 Análisis Bivariado

**Edad:**

La variable Edad se distribuye en tres rangos de [18 - 42] que representa el 9%, de [43 - 60] que representa el 60% y el rango de [61 - mas] que representa el 31%. (Ver Anexos Tabla N° 4.22)

Grafico N° 4.1
EDAD VS VARIABLE RESPUESTA

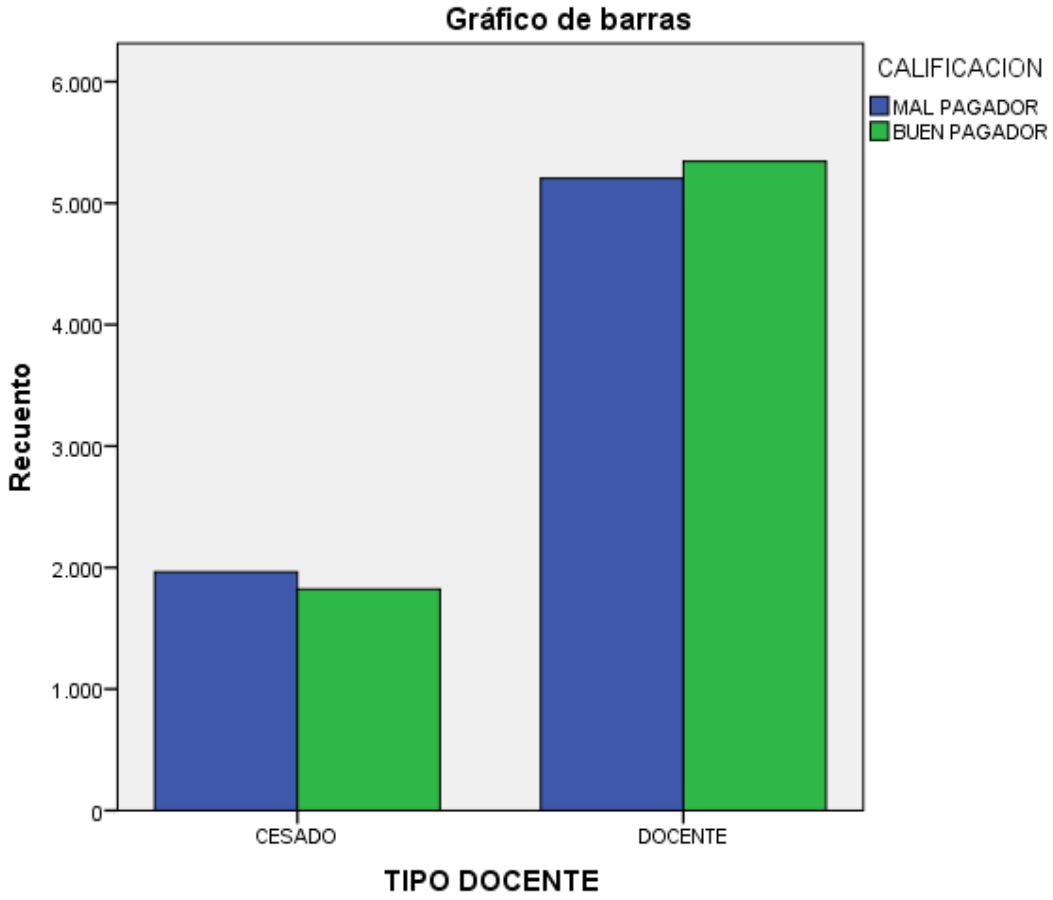


**Tipo de docente:**

La variable Tipo de Docente se divide en Cesado y Docente Activo que representan 26% y 74%, respectivamente. (Ver Anexos Tabla N° 4.23)

Grafico N° 4.2

TIPO DOCENTE VS VARIABLE RESPUESTA

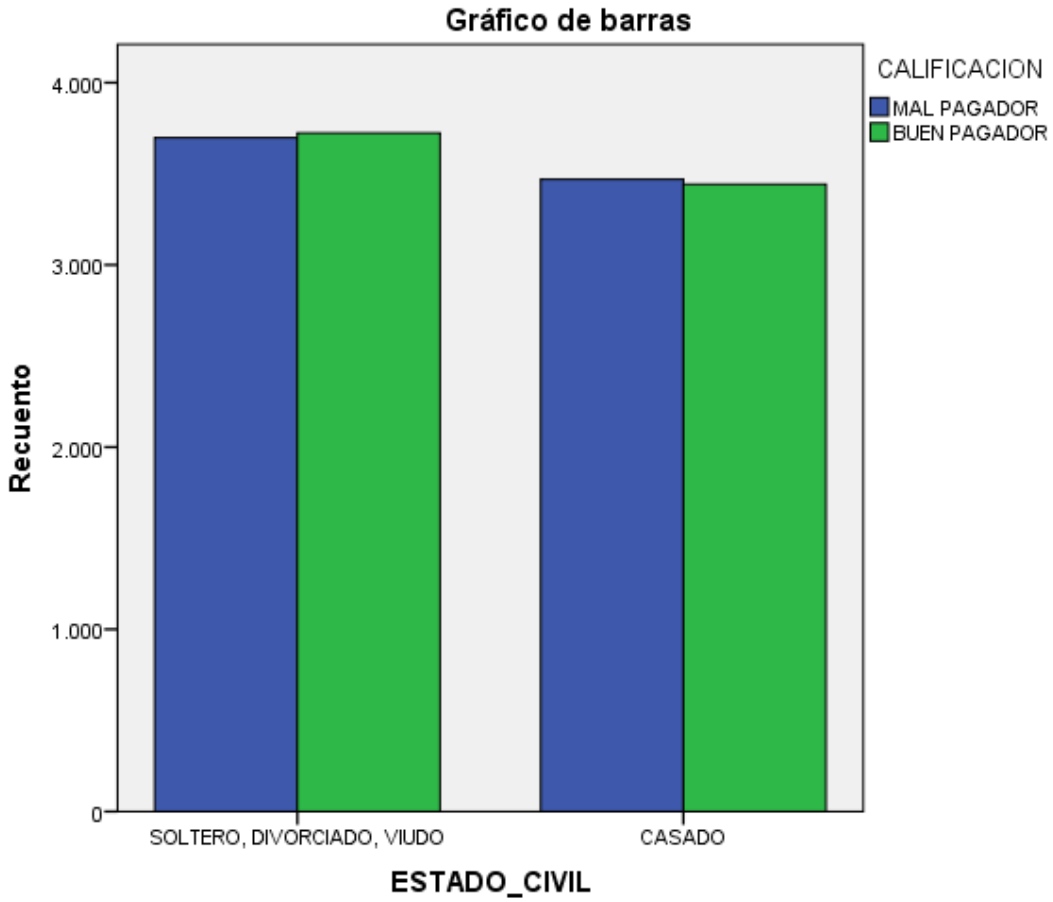


**Estado Civil:**

La variable Estado Civil ha sido dividida en dos categorías: Casado y Soltero, Divorciado, Viudo las cuales representan 48% y 52%, para cada categoría. (Ver Anexos Tabla N° 4.24)

Grafico N° 4.3

ESTADO CIVIL VS VARIABLE RESPUESTA

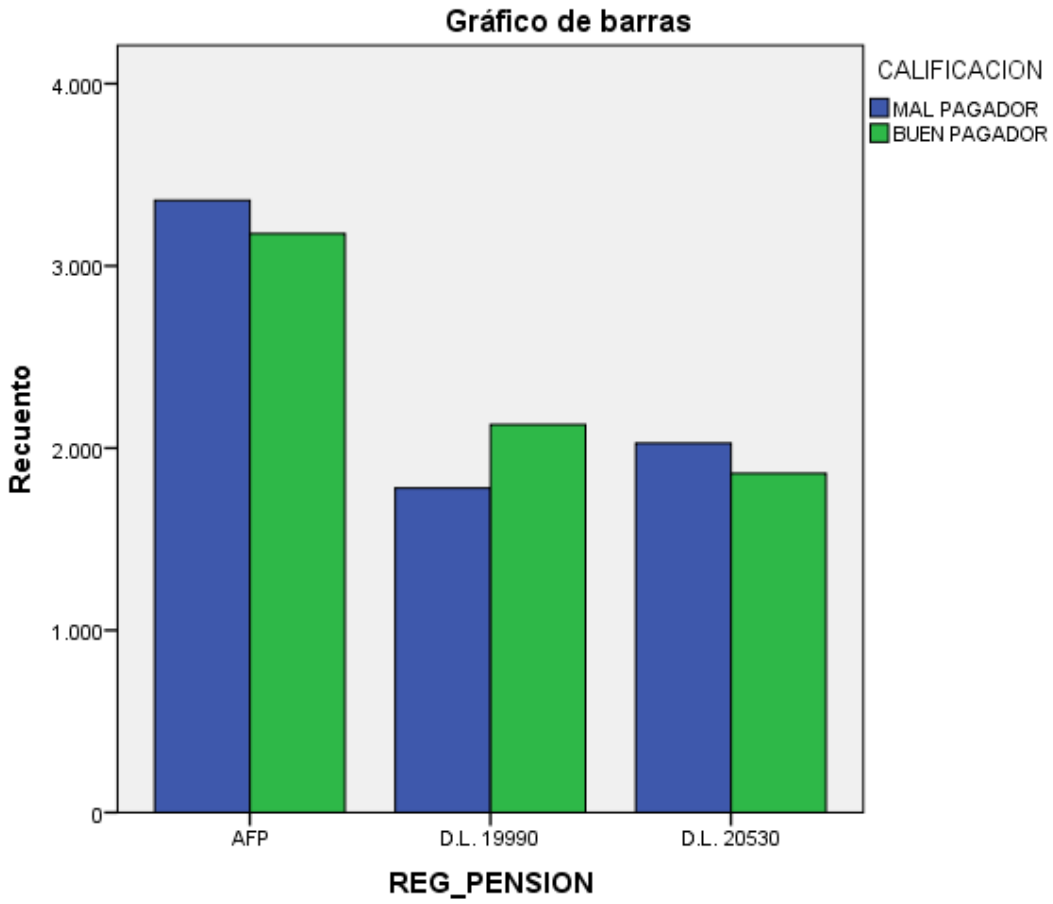


**Régimen de Pensión:**

La variable Régimen de Pensión consta de tres categorías, la categoría AFP que representa un 46% de la data, la categoría D.L. 19990 representa 27% de la data y D.L. 20530 que representa el 27% restante de la data.(Ver Anexos Tabla N° 4.25)

Grafico N° 4.4

REGIMEN PENSION VS VARIABLE RESPUESTA

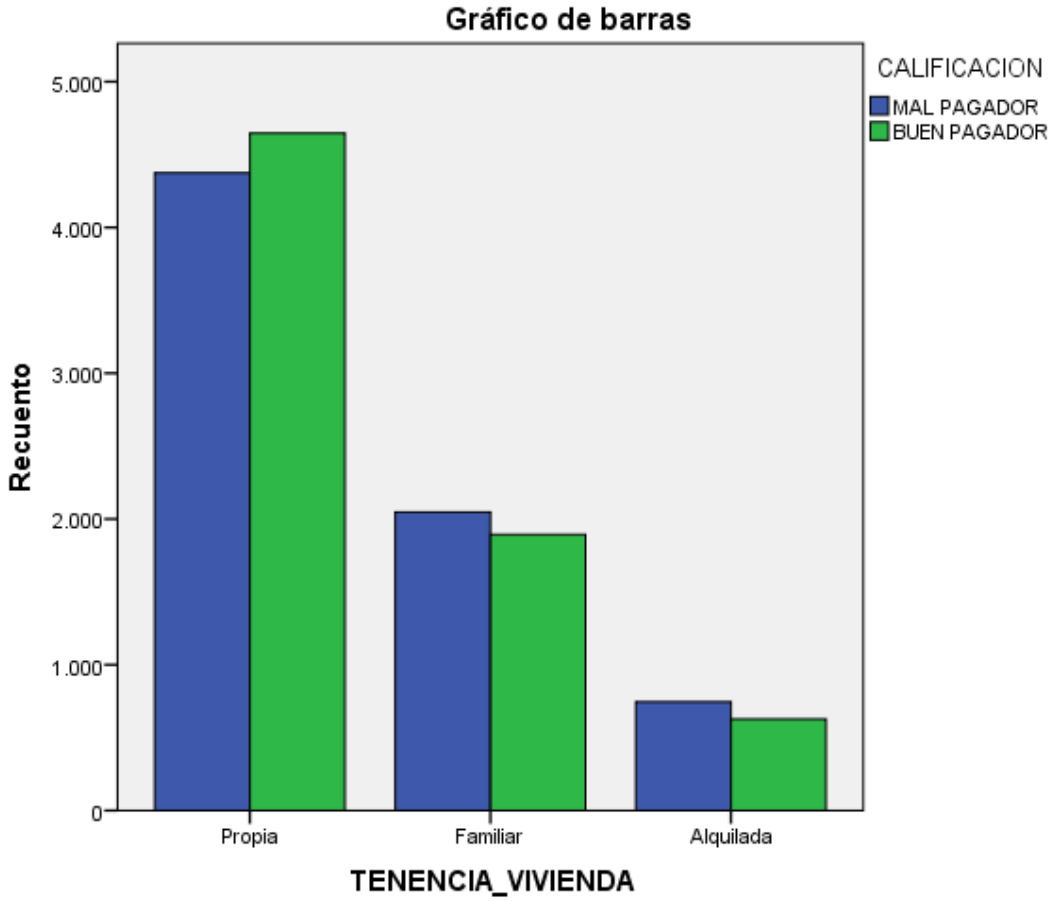


**Tenencia de Vivienda:**

La variable Tenencia de Vivienda se divide en tres categorías: Propia, Familiar y Alquilada que representan el 63%, 27% y 10% respectivamente.(Ver Anexos Tabla N° 4.26)

Grafico N° 4.5

TENENCIA DE VIVIENDA VS VARIABLE RESPUESTA

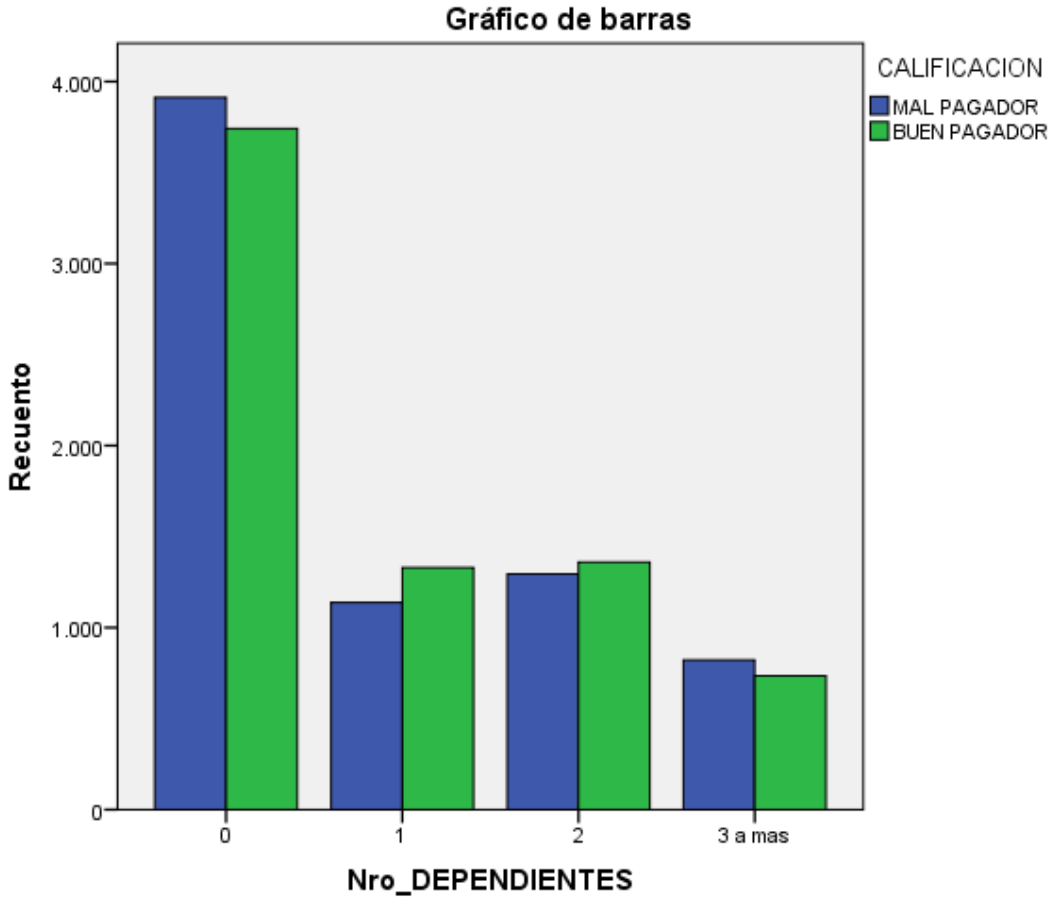


**N° Dependientes:**

La variable Número de Dependientes se dividió la data en los clientes que tienen 0, 1, 2 y más de 3 dependientes. El número de dependientes es el número de personas que dependen económicamente del cliente como son los hijos por ejemplo. El porcentaje de clientes que no tiene dependientes es del 53%, el 9% de clientes tiene un dependiente, los clientes que tienen dos dependientes representan 9% y los que tienen tres a más dependientes representan el 11%.(Ver Anexos Tabla N° 4.27)

Grafico N° 4.6

NUMERO DE DEPENDIENTES VS VARIABLE RESPUESTA

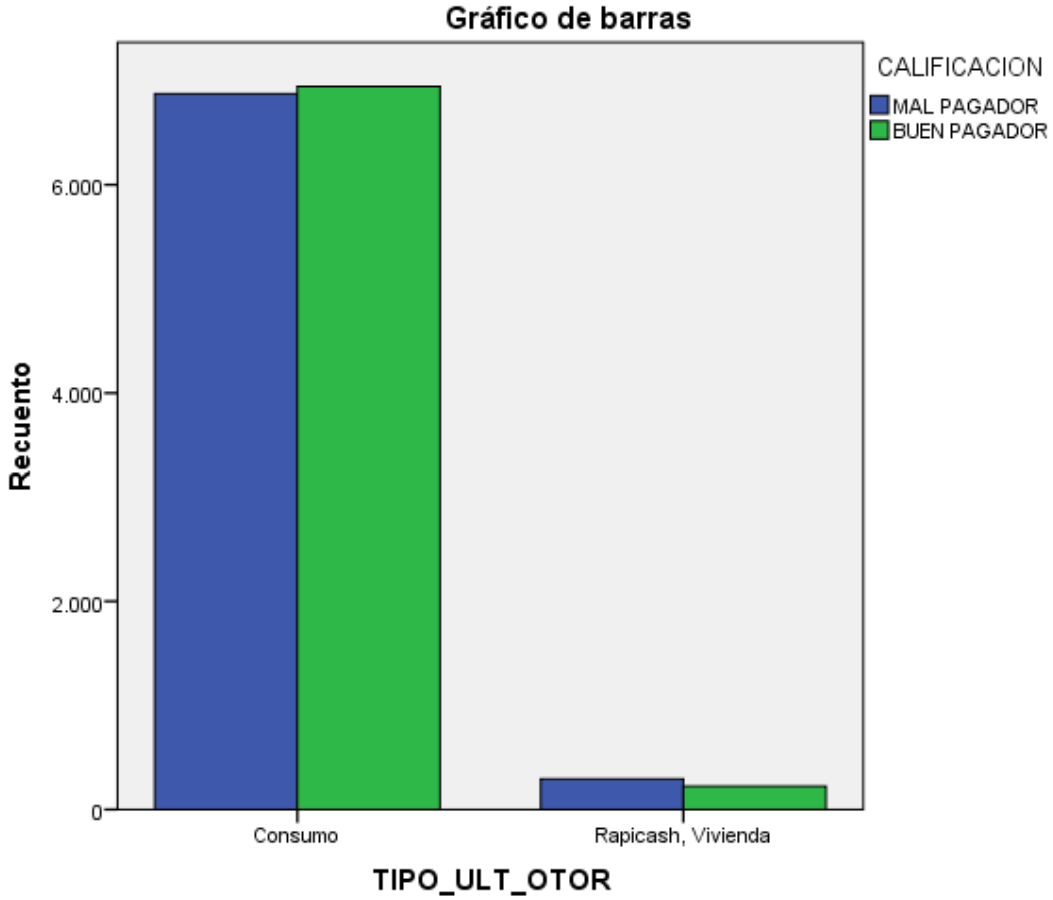


**Tipo de Último Crédito:**

En la Entidad Bancaria existen tipos de créditos los cuales han sido agrupados en dos categorías: Crédito Consumo que representa el 96% del total, la otra categoría (Rapicash y Vivienda) representa el 4%.(Ver Anexos Tabla N° 4.28)

Grafico N° 4.7

TIPO ÚLTIMO CREDITO VS VARIABLE RESPUESTA

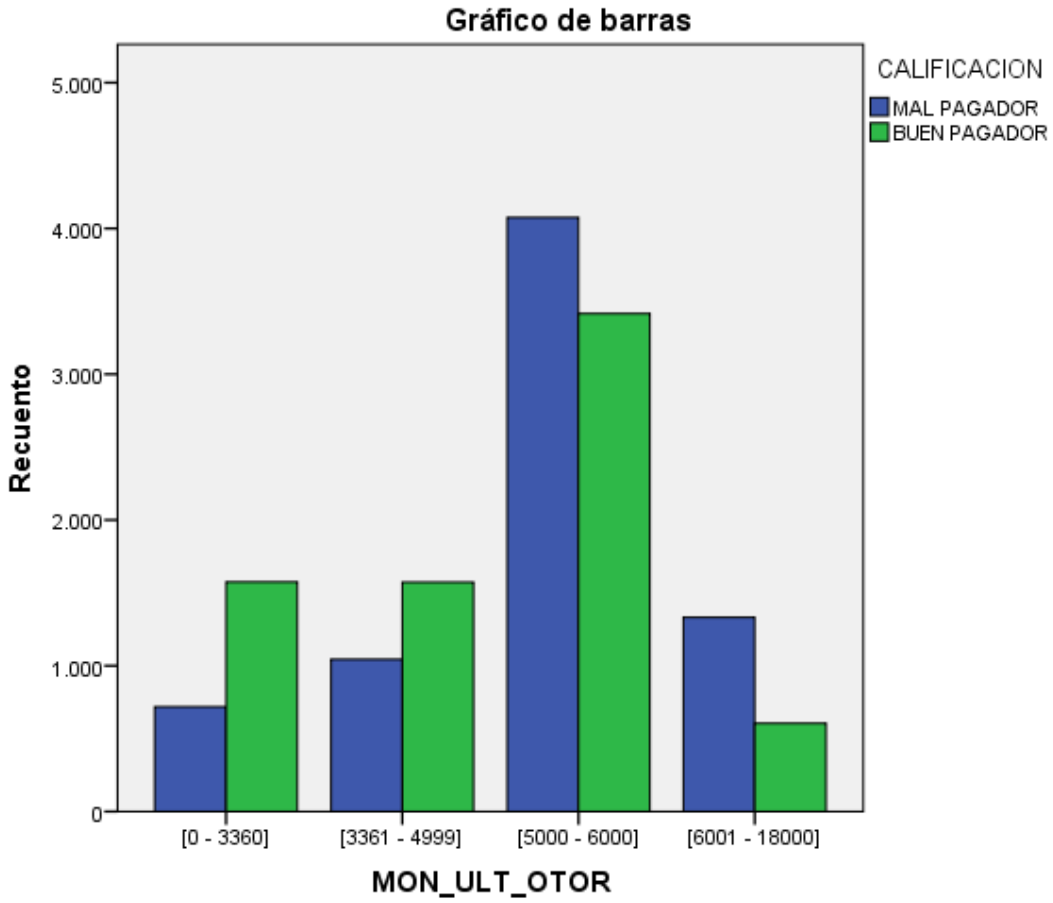


**Monto de Último Crédito:**

Los clientes han solicitado en alguna ocasión algún crédito, en esta variable se forman los rangos de los montos de los créditos. El primer rango [0 - 3360]representa el 16%, el segundo rango [3361 - 4999] representa el 18%, el tercer rango [5000 - 6000] representa el 52% del total y el cuarto rango [6001 - 18000] representa el 13%. (Ver Anexos Tabla N° 4.29)

Grafico N° 4.8

MONTO ÚLTIMO CREDITO VS VARIABLE RESPUESTA

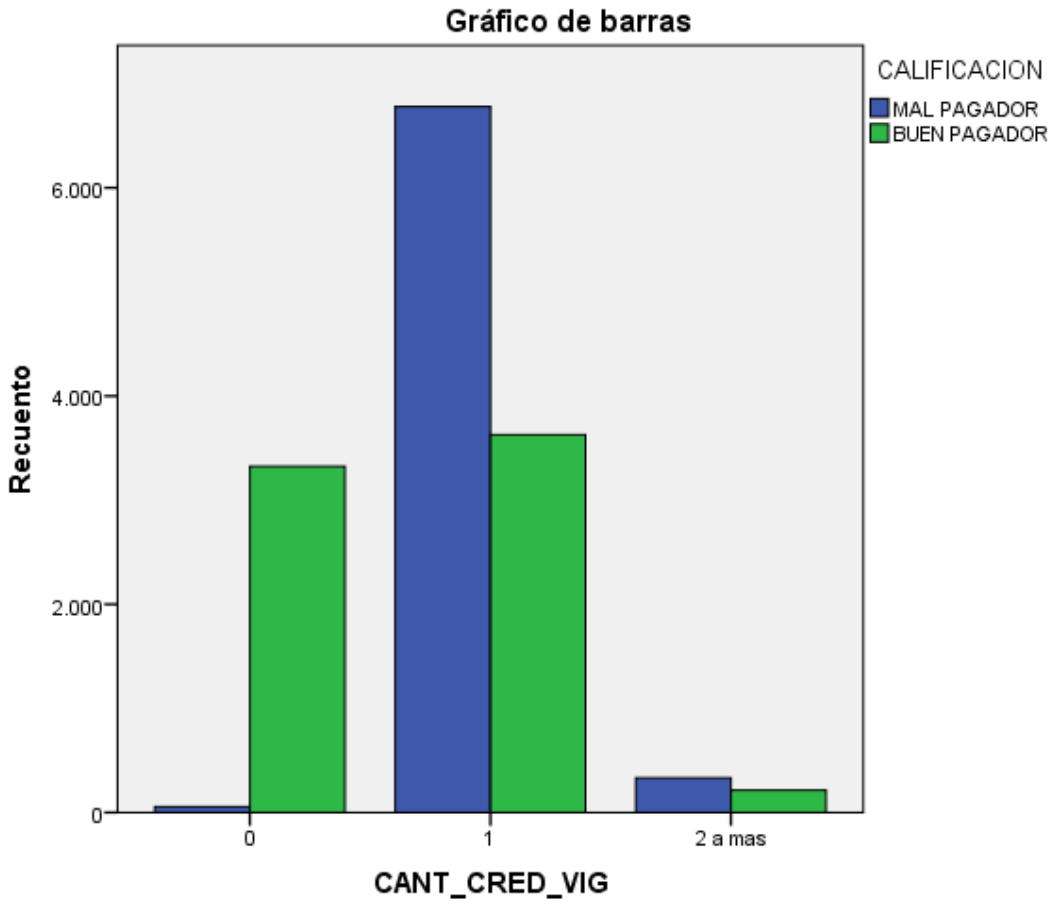


**Cantidad de Créditos Vigentes:**

Al momento del análisis los clientes tenían créditos vigentes. Los que no tenían créditos vigentes representan el 24% del total, los que tenían un crédito vigente representan el 73% y el 4% restante tenía dos o más créditos vigentes.(Ver Anexos Tabla N° 4.30)

Grafico N° 4.9

CANTIDAD CREDITOS VIGENTES VS VARIABLE RESPUESTA

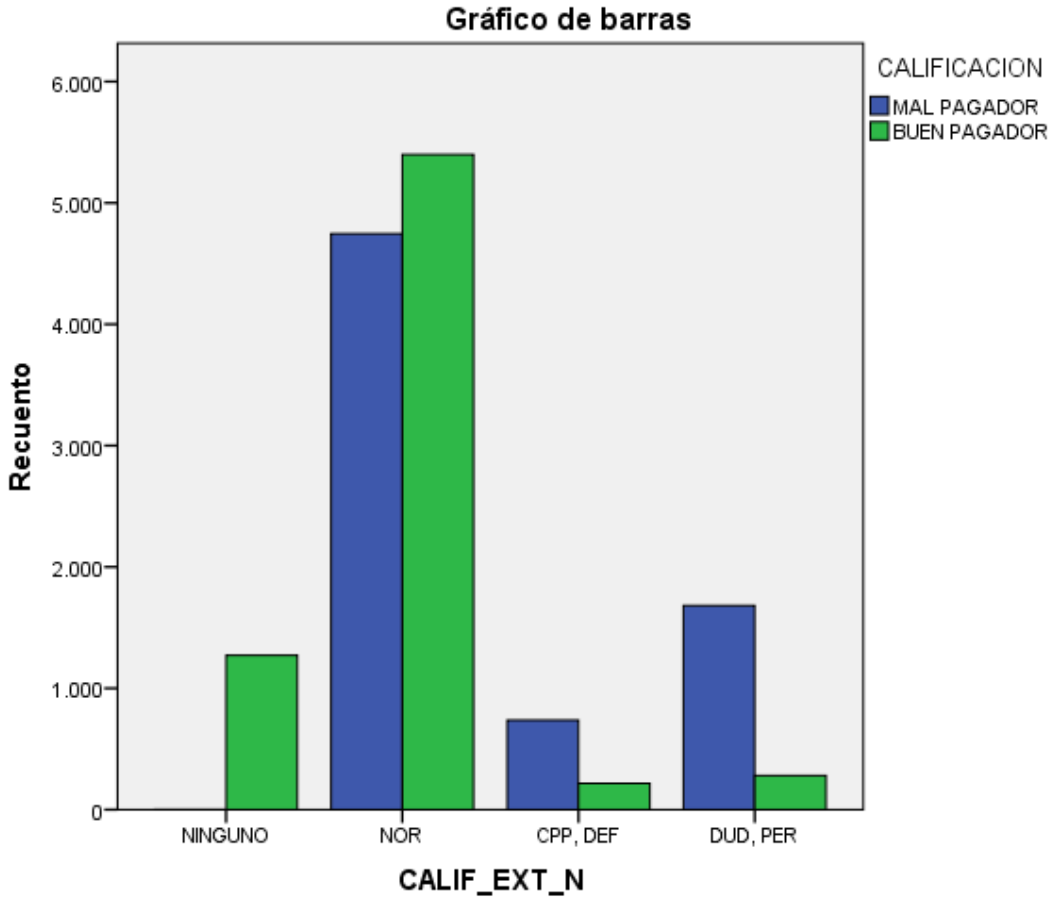


**Calificación Externa:**

La calificación Externa que reporta la Superintendencia de Banca, Seguros y AFP'sa través del Reporte Consolidados de Clientes (RCC), Existen Clientes que no tienen calificación Externa la cual representa el 9% del total, los clientes que tienen calificación Normal representan el 71%, las calificaciones CPP y Deficiente representan el 7% y el 14% restante tienen las calificaciones de Dudoso y Perdida.(Ver Anexos Tabla N° 4.32)

Grafico N° 4.11

CALIFICACION EXTERNA VS VARIABLE RESPUESTA

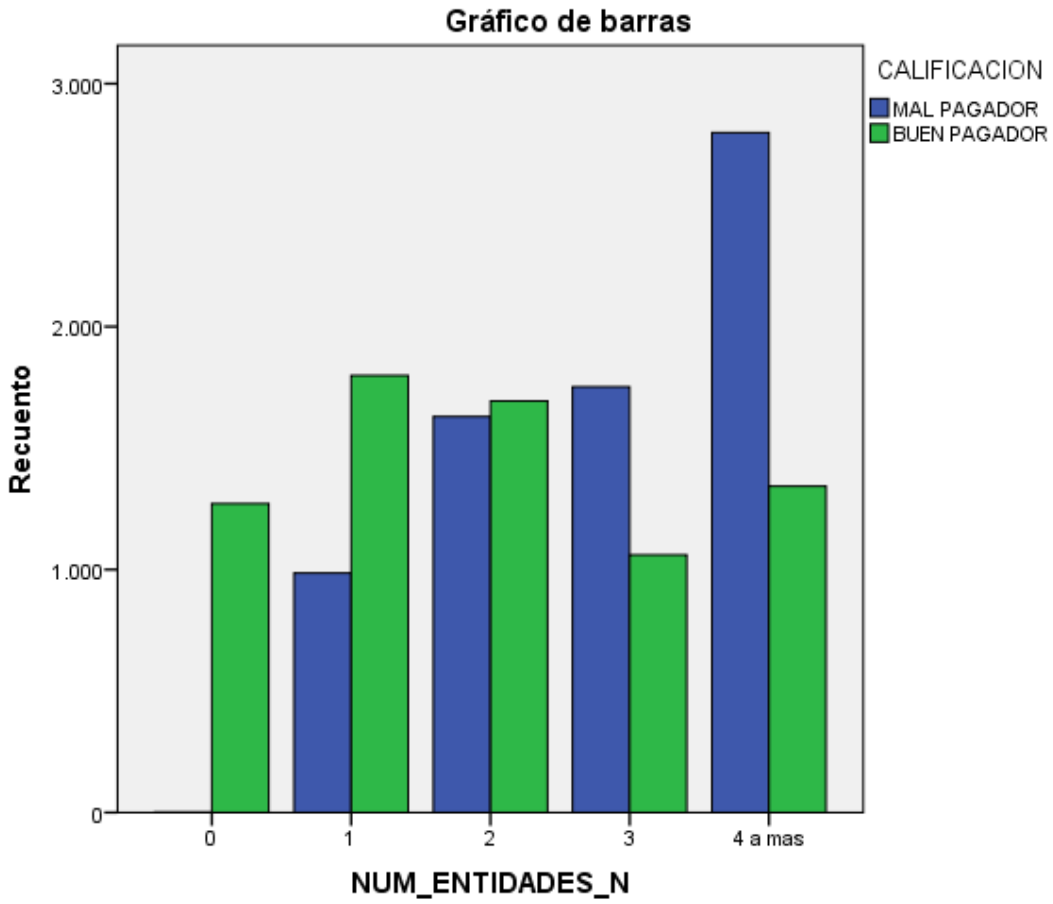


**Numero Entidades:**

Los clientes que al momento del análisis no tenían créditos en ninguna Entidad es el 9% de la data, los clientes que tienen una entidad es el 19%, con respecto a los clientes que tienen 2, 3 y 4 a más son 23%, 20% y 29%, respectivamente.(Ver Anexos Tabla N° 4.33)

Grafico N° 4.12

NUMERO ENTIDADES VS VARIABLE RESPUESTA

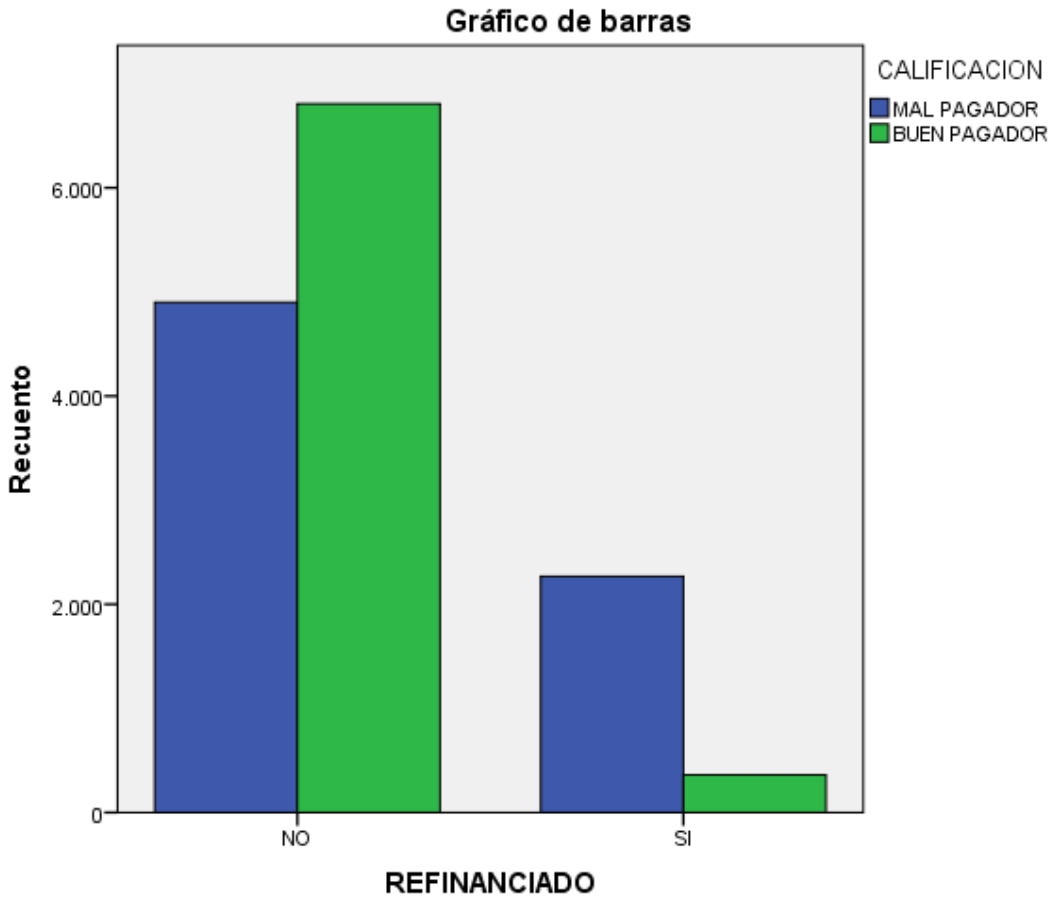


**Refinanciado:**

Los clientes que no han tenido un crédito Refinanciado representan el 82% y el 18% restante si ha tenido en alguna oportunidad algún crédito refinanciado.(Ver Anexos Tabla N° 4.34)

Grafico N° 4.13

REFINANCIADO VS VARIABLE RESPUESTA

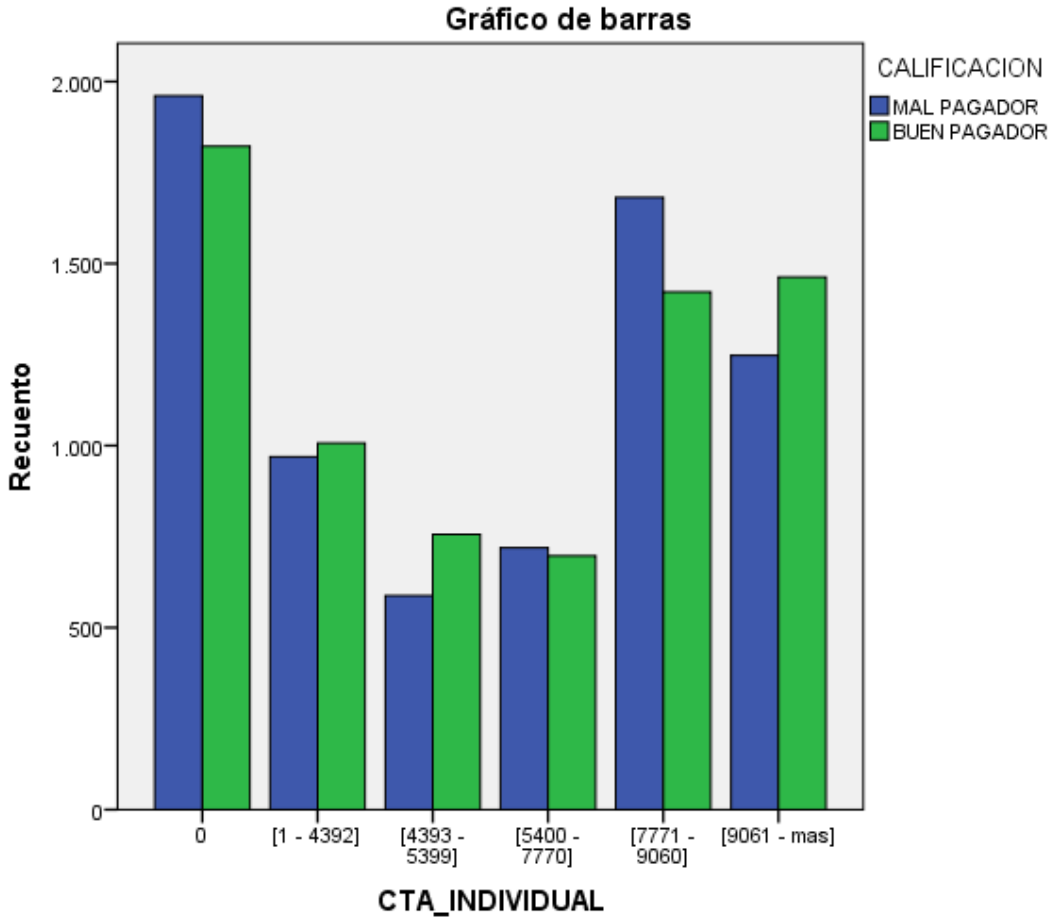


**Cuenta Individual:**

La cuenta individual se refiere a los aportes de los docentes. Los rangos del monto en su cuenta individual son seis: El que no cuenta con Saldo en su cuenta Individual representa el 26% de los clientes, los que tienen un saldo menor a 4,392 soles representan el 14%, el 9% pertenecen al rango [4393 - 5399], el 10% pertenecen al rango [5400 - 7770], el 22% pertenecen al rango [7771 - 9060] y el 19% restante pertenecen al rango [9061 - mas].(Ver Anexos Tabla N° 4.35)

Grafico N° 4.14

CUENTA INDIVIDUAL VS VARIABLE RESPUESTA

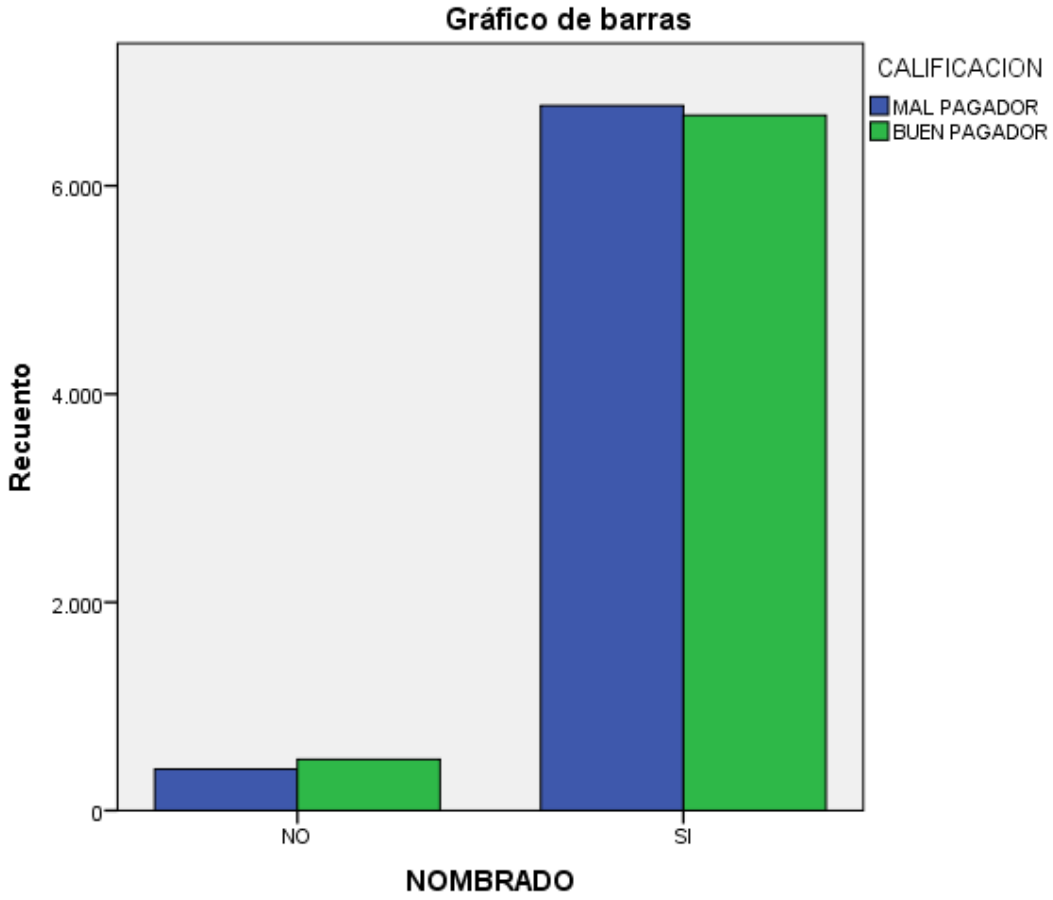


**Nombrado:**

Los clientes pueden ser Nombrados por el Estado o ser Contratados, el 94% son Nombrados ye el 6% restante son contratados.(Ver Anexos Tabla N° 4.36)

Grafico N° 4.15

NOMBRADO VS VARIABLE RESPUESTA

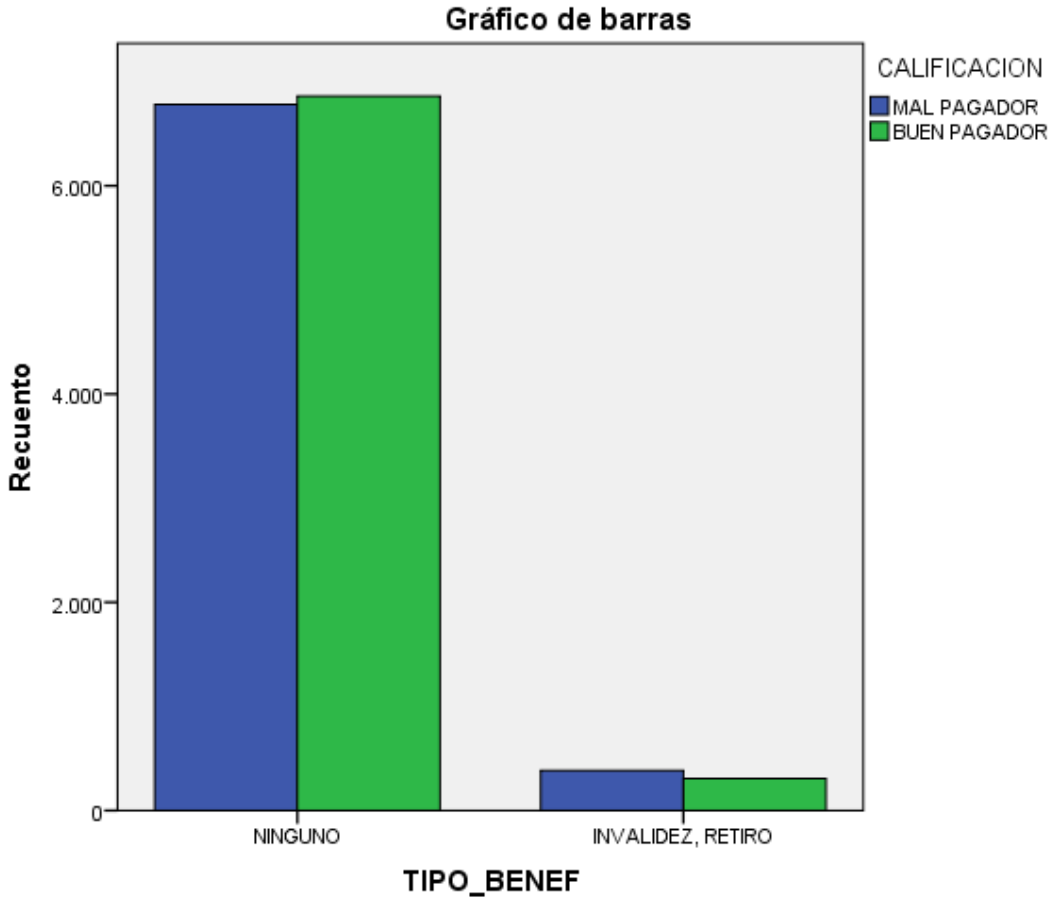


**Tipo Beneficio:**

Al momento de su pase a retiro del Sector Magisterial el docente puede recibir sus beneficios por Retiro, el 95% de la data no han recibido su beneficio por retiro y el 5% restante si han recibido el beneficio sea por retiro o invalidez.(Ver Anexos Tabla N° 4.37)

Grafico N° 4.16

TIPO BENEFICIO VS VARIABLE RESPUESTA

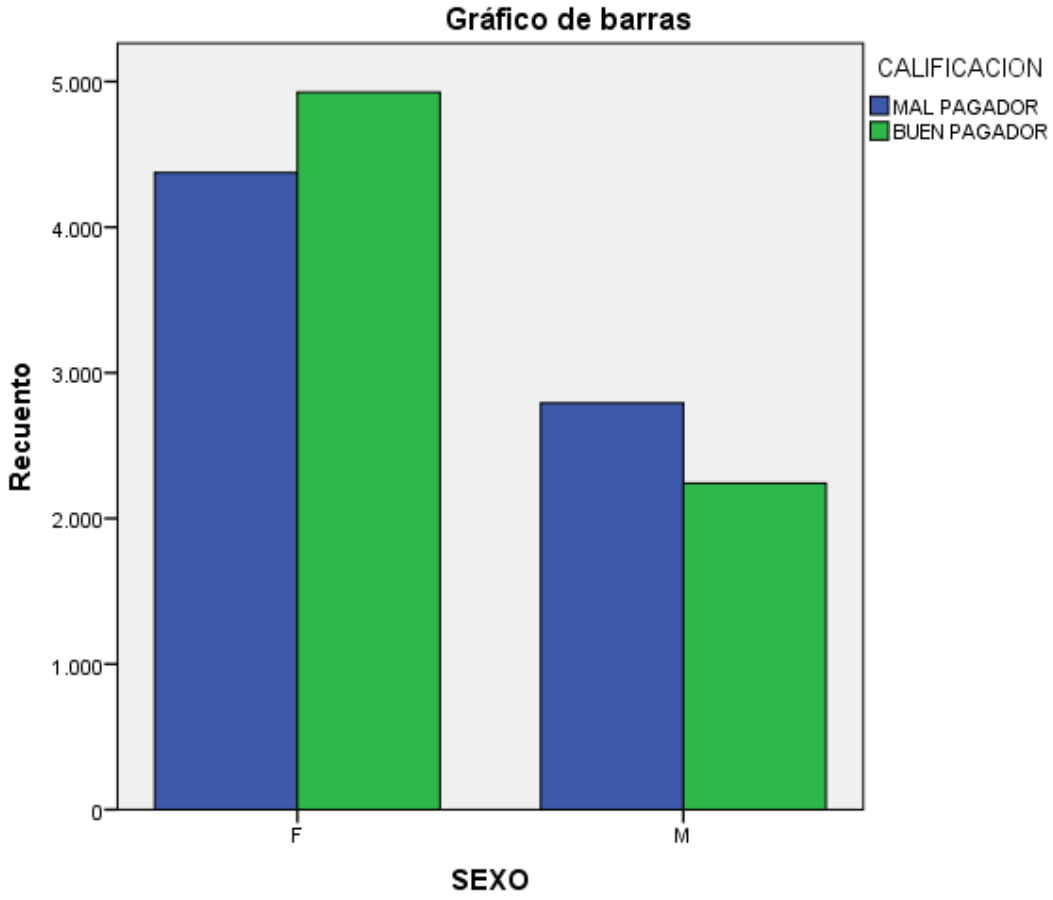


**Género:**

Con respecto al género el 65% de la data corresponde a las mujeres y el 35% a los docentes varones.(Ver Anexos Tabla N° 4.38)

Grafico N° 4.17

GENERO VS VARIABLE RESPUESTA

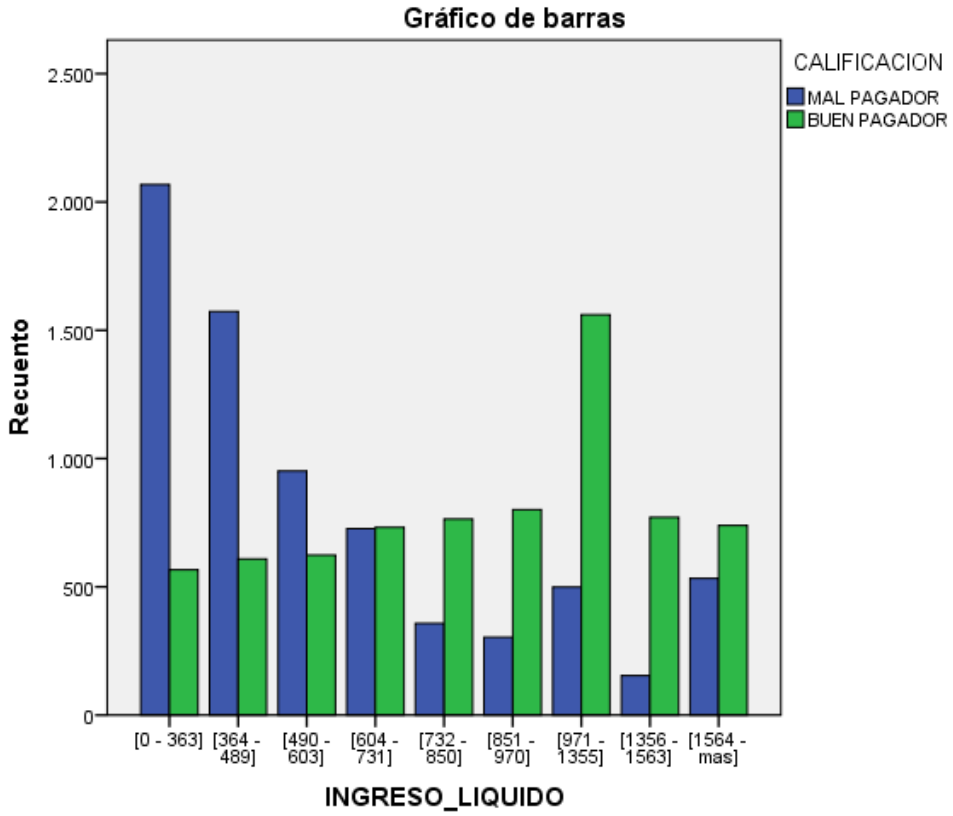


**Ingreso Líquido:**

Los clientes del análisis tienen un nivel de ingresos el cual ha sido dividido en 9 rangos: el primer rango [0 - 363] representa el 18%, el segundo rango [364 - 489] representa el 15%, el tercer rango [490 - 603] representa el 11%, el cuarto rango [604 - 731] representa el 10%, el quinto rango [732 - 850] representa el 8%, el sexto rango [851 - 970] representa el 8%, el séptimo rango [971 - 1355] representa el 14%, el octavo rango [1356 - 1563] representa el 6%, el noveno rango [1564 - mas] representa el 9%.(Ver Anexos Tabla N° 4.39)

Grafico N° 4.18

INGRESO LIQUIDO VS VARIABLE RESPUESTA

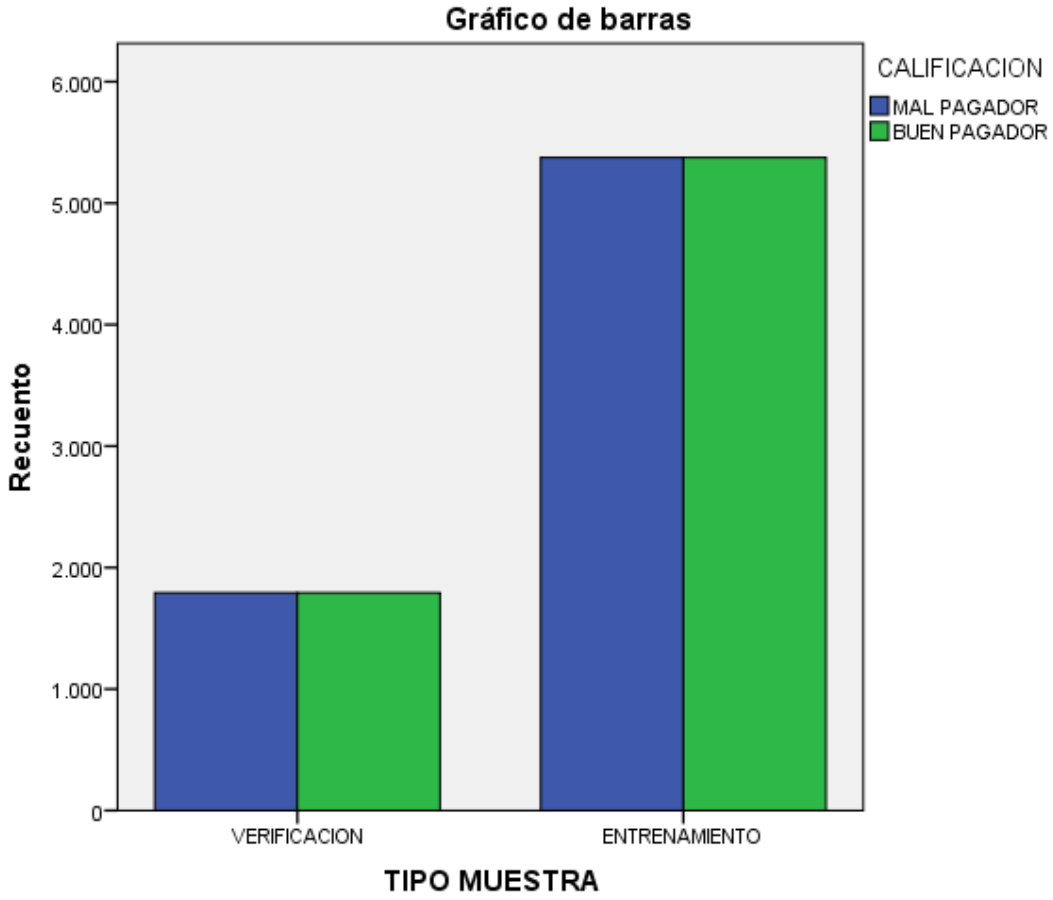


**Tipo Muestra:**

La data se divide en dos tipos de muestras de Entrenamiento que tiene el 75% de la data y de Verificación que tiene el 25% restante.(Ver Anexos Tabla N° 4.40)

Grafico N° 4.19

TIPO MUESTRA VS VARIABLE RESPUESTA



## 4.2 Construcción del Modelo Logístico

Luego de haber definido las variables candidatas y de haber realizado el análisis univariado y bivariado de las variables. Se procede a construir el modelo logístico mediante el método de máxima verosimilitud.

En el presente estudio la cantidad de buenos asciende al 88% del total de registros y el número de clientes malos asciende a 12% se decide tomar el total de la población debido a que cuentan con la información en las variables candidatas.

**Variable Respuesta Calificación:**

Tabla N° 4.2.1
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE RESPUESTA

| CALIFICACION | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | MAL PAGADOR | 7,167 | 50.0% | 50.0% | 50.0% |
| | BUEN PAGADOR | 7,167 | 50.0% | 50.0% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

Se particiona la muestra en dos sub muestras, una la *muestra de entrenamiento* la cual servirá para desarrollar el modelo que comprende el 75% de la muestra y la segunda llamada *muestra de verificación*para la validación del modelo que representa el 25% de la muestra, sin dejar de lado las proporciones de buenos y malos.

**Tipo de Muestra:**

Tabla N° 4.2.2
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE RESPUESTA

| TIPO MUESTRA | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | VERIFICACION | 3,584 | 25.0% | 25.0% | 25.0% |
| | ENTRENAMIENTO | 10,750 | 75.0% | 75.0% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

Para la construcción del modelo se utilizó el método Hacia Atrás (Wald), para evaluar la significancia del modelo logístico se observa los parámetros (B) y su significancia (Sig) debe ser menor de 0.05 entonces la variable explica si un cliente podría ser bueno o malo, en caso contrario la variable debe ser eliminada del modelo.

Tabla N° 4.2.3
VARIABLES EN LA ECUACIÓN

| Variable en la Ecuación | Etiqueta | B | Error estándar | Wald | gl | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| RANGOEDAD | [18 - 42] | | | 8.139 | 2 | .017 | |
| RANGOEDAD(1) | [43 - 60] | .281 | .100 | 7.848 | 1 | .005 | 1.324 |
| RANGOEDAD(2) | [61 - mas] | .221 | .139 | 2.535 | 1 | .111 | 1.247 |
| TENENCIA_VIVIENDA | Propia | | | 5.528 | 2 | .063 | |
| TENENCIA_VIVIENDA(1) | Familiar | .067 | .059 | 1.305 | 1 | .253 | 1.070 |
| TENENCIA_VIVIENDA(2) | Alquilada | -.158 | .090 | 3.114 | 1 | .078 | .854 |
| Nro_DEPENDIENTES | 0 | | | 6.790 | 3 | .079 | |
| Nro_DEPENDIENTES(1) | 1 | .110 | .074 | 2.256 | 1 | .133 | 1.117 |
| Nro_DEPENDIENTES(2) | 2 | .056 | .073 | .578 | 1 | .447 | 1.057 |
| Nro_DEPENDIENTES(3) | 3 a mas | -.144 | .089 | 2.611 | 1 | .106 | .866 |
| TIPO_ULT_OTOR | Consumo | | | | | .000 | |
| TIPO_ULT_OTOR(1) | Rapicash, Vivienda | -.236 | .142 | 2.778 | 1 | .096 | .789 |
| MON_ULT_OTOR | [0 - 3360] | | | 13.589 | 3 | .004 | |
| MON_ULT_OTOR(1) | [3361 - 4999] | -.058 | .105 | .306 | 1 | .580 | .943 |
| MON_ULT_OTOR(2) | [5000 - 6000] | -.184 | .089 | 4.314 | 1 | .038 | .832 |
| MON_ULT_OTOR(3) | [6001 - 18000] | -.362 | .110 | 10.783 | 1 | .001 | .697 |
| CANT_CRED_VIG | 0 | | | 533.751 | 2 | .000 | |
| CANT_CRED_VIG(1) | 1 | -3.861 | .167 | 533.279 | 1 | .000 | .021 |
| CANT_CRED_VIG(2) | 2 a mas | -3.807 | .198 | 368.286 | 1 | .000 | .022 |
| CALIF_EXT_N | NINGUNO | | | 352.502 | 3 | .000 | |
| CALIF_EXT_N(1) | NOR | -4.034 | 1.014 | 15.841 | 1 | .000 | .018 |
| CALIF_EXT_N(2) | CPP, DEF | -4.887 | 1.019 | 23.025 | 1 | .000 | .008 |
| CALIF_EXT_N(3) | DUD, PER | -5.701 | 1.017 | 31.424 | 1 | .000 | .003 |
| NUM_ENTIDADES_N | 0 | | | 27.205 | 3 | .000 | |
| NUM_ENTIDADES_N(1) | 1 | .241 | .078 | 9.412 | 1 | .002 | 1.272 |
| NUM_ENTIDADES_N(2) | 2 | .317 | .068 | 21.452 | 1 | .000 | 1.373 |
| NUM_ENTIDADES_N(3) | 3 | .033 | .071 | .225 | 1 | .635 | 1.034 |
| NUM_ENTIDADES_N(4) | 4 a mas | | | | | | |
| REFINANCIADO | NO | | | | | | |
| REFINANCIADO(1) | SI | -.954 | .082 | 136.875 | 1 | .000 | .385 |
| CTA_INDIVIDUAL | 0 | | | 64.679 | 5 | .000 | |
| CTA_INDIVIDUAL(1) | [1 - 4392] | .178 | .137 | 1.694 | 1 | .193 | 1.195 |
| CTA_INDIVIDUAL(2) | [4393 - 5399] | .288 | .139 | 4.267 | 1 | .039 | 1.333 |
| CTA_INDIVIDUAL(3) | [5400 - 7770] | -.203 | .141 | 2.066 | 1 | .151 | .816 |
| CTA_INDIVIDUAL(4) | [7771 - 9060] | -.314 | .122 | 6.616 | 1 | .010 | .730 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CTA_INDIVIDUAL(5) | [9061 - mas] | .202 | .112 | 3.266 | 1 | .071 | 1.224 |
| TIPO_BENEF | NINGUNO | | | | | | |
| TIPO_BENEF(1) | INVALIDEZ, RETIRO | -.428 | .139 | 9.493 | 1 | .002 | .652 |
| SEXO | F | | | | | | |
| SEXO(1) | M | -.336 | .055 | 37.144 | 1 | .000 | .715 |
| INGRESO_LIQUIDO | [0 - 363] | | | 344.204 | 8 | .000 | |
| INGRESO_LIQUIDO(1) | [364 - 489] | .188 | .093 | 4.065 | 1 | .044 | 1.207 |
| INGRESO_LIQUIDO(2) | [490 - 603] | .741 | .097 | 58.412 | 1 | .000 | 2.098 |
| INGRESO_LIQUIDO(3) | [604 - 731] | .875 | .099 | 77.317 | 1 | .000 | 2.398 |
| INGRESO_LIQUIDO(4) | [732 - 850] | 1.317 | .111 | 139.482 | 1 | .000 | 3.730 |
| INGRESO_LIQUIDO(5) | [851 - 970] | 1.423 | .117 | 148.297 | 1 | .000 | 4.151 |
| INGRESO_LIQUIDO(6) | [971 - 1355] | 1.340 | .105 | 161.991 | 1 | .000 | 3.819 |
| INGRESO_LIQUIDO(7) | [1356 - 1563] | 1.005 | .154 | 42.404 | 1 | .000 | 2.732 |
| INGRESO_LIQUIDO(8) | [1564 - mas] | .316 | .122 | 6.746 | 1 | .009 | 1.372 |
| Constante | | 6.952 | 1.031 | 45.474 | 1 | .000 | 1045.416 |

FUENTE: Entidad Bancaria

ELABORACIÓN: Propia

Solamente se decide incluir en el modelo a las variables que por lo menos alguna de sus categorías tiene significancia (Sig) menor 0.05. Las demás serán eliminadas del modelo. En el siguiente cuadro se presentan las variables incluidas en el modelo.

Tabla N° 4.2.4

VARIABLES EN LA ECUACIÓN

| Variable en la Ecuación | Etiqueta | B | Error estándar | Wald | gl | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| RANGOEDAD | [18 - 42] | | | 8.139 | 2 | .017 | |
| RANGOEDAD(1) | [43 - 60] | .281 | .100 | 7.848 | 1 | .005 | 1.324 |
| RANGOEDAD(2) | [61 - mas] | .221 | .139 | 2.535 | 1 | .111 | 1.247 |
| TENENCIA_VIVIENDA | Propia | | | 5.528 | 2 | .063 | |
| TENENCIA_VIVIENDA(1) | Familiar | .067 | .059 | 1.305 | 1 | .253 | 1.070 |
| TENENCIA_VIVIENDA(2) | Alquilada | -.158 | .090 | 3.114 | 1 | .078 | .854 |
| Nro_DEPENDIENTES | 0 | | | 6.790 | 3 | .079 | |
| Nro_DEPENDIENTES(1) | 1 | .110 | .074 | 2.256 | 1 | .133 | 1.117 |
| Nro_DEPENDIENTES(2) | 2 | .056 | .073 | .578 | 1 | .447 | 1.057 |
| Nro_DEPENDIENTES(3) | 3 a mas | -.144 | .089 | 2.611 | 1 | .106 | .866 |
| TIPO_ULT_OTOR | Consumo | | | | | .000 | |
| TIPO_ULT_OTOR(1) | Rapicash, Vivienda | -.236 | .142 | 2.778 | 1 | .096 | .789 |
| MON_ULT_OTOR | [0 - 3360] | | | 13.589 | 3 | .004 | |
| MON_ULT_OTOR(1) | [3361 - 4999] | -.058 | .105 | .306 | 1 | .580 | .943 |
| MON_ULT_OTOR(2) | [5000 - 6000] | -.184 | .089 | 4.314 | 1 | .038 | .832 |
| MON_ULT_OTOR(3) | [6001 - 18000] | -.362 | .110 | 10.783 | 1 | .001 | .697 |
| CANT_CRED_VIG | 0 | | | 533.751 | 2 | .000 | |
| CANT_CRED_VIG(1) | 1 | -3.861 | .167 | 533.279 | 1 | .000 | .021 |
| CANT_CRED_VIG(2) | 2 a mas | -3.807 | .198 | 368.286 | 1 | .000 | .022 |
| CALIF_EXT_N | NINGUNO | | | 352.502 | 3 | .000 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CALIF_EXT_N(1) | NOR | -4.034 | 1.014 | 15.841 | 1 | .000 | .018 |
| CALIF_EXT_N(2) | CPP, DEF | -4.887 | 1.019 | 23.025 | 1 | .000 | .008 |
| CALIF_EXT_N(3) | DUD, PER | -5.701 | 1.017 | 31.424 | 1 | .000 | .003 |
| NUM_ENTIDADES_N | 0 | | | 27.205 | 3 | .000 | |
| NUM_ENTIDADES_N(1) | 1 | .241 | .078 | 9.412 | 1 | .002 | 1.272 |
| NUM_ENTIDADES_N(2) | 2 | .317 | .068 | 21.452 | 1 | .000 | 1.373 |
| NUM_ENTIDADES_N(3) | 3 | .033 | .071 | .225 | 1 | .635 | 1.034 |
| NUM_ENTIDADES_N(4) | 4 a mas | | | | | | |
| REFINANCIADO | NO | | | | | | |
| REFINANCIADO(1) | SI | -.954 | .082 | 136.875 | 1 | .000 | .385 |
| CTA_INDIVIDUAL | 0 | | | 64.679 | 5 | .000 | |
| CTA_INDIVIDUAL(1) | [1 - 4392] | .178 | .137 | 1.694 | 1 | .193 | 1.195 |
| CTA_INDIVIDUAL(2) | [4393 - 5399] | .288 | .139 | 4.267 | 1 | .039 | 1.333 |
| CTA_INDIVIDUAL(3) | [5400 - 7770] | -.203 | .141 | 2.066 | 1 | .151 | .816 |
| CTA_INDIVIDUAL(4) | [7771 - 9060] | -.314 | .122 | 6.616 | 1 | .010 | .730 |
| CTA_INDIVIDUAL(5) | [9061 - mas] | .202 | .112 | 3.266 | 1 | .071 | 1.224 |
| TIPO_BENEF | NINGUNO | | | | | | |
| TIPO_BENEF(1) | INVALIDEZ, RETIRO | -.428 | .139 | 9.493 | 1 | .002 | .652 |
| SEXO | F | | | | | | |
| SEXO(1) | M | -.336 | .055 | 37.144 | 1 | .000 | .715 |
| INGRESO_LIQUIDO | [0 - 363] | | | 344.204 | 8 | .000 | |
| INGRESO_LIQUIDO(1) | [364 - 489] | .188 | .093 | 4.065 | 1 | .044 | 1.207 |
| INGRESO_LIQUIDO(2) | [490 - 603] | .741 | .097 | 58.412 | 1 | .000 | 2.098 |
| INGRESO_LIQUIDO(3) | [604 - 731] | .875 | .099 | 77.317 | 1 | .000 | 2.398 |
| INGRESO_LIQUIDO(4) | [732 - 850] | 1.317 | .111 | 139.482 | 1 | .000 | 3.730 |
| INGRESO_LIQUIDO(5) | [851 - 970] | 1.423 | .117 | 148.297 | 1 | .000 | 4.151 |
| INGRESO_LIQUIDO(6) | [971 - 1355] | 1.340 | .105 | 161.991 | 1 | .000 | 3.819 |
| INGRESO_LIQUIDO(7) | [1356 - 1563] | 1.005 | .154 | 42.404 | 1 | .000 | 2.732 |
| INGRESO_LIQUIDO(8) | [1564 - mas] | .316 | .122 | 6.746 | 1 | .009 | 1.372 |
| Constante | | 6.952 | 1.031 | 45.474 | 1 | .000 | 1045.416 |

FUENTE: Entidad Bancaria

ELABORACIÓN: Propia

Las variables que pueden explicar si un cliente va ser un buen pagador o un mal pagador son: Edad, Numero de Dependientes, Calificación Externa, Número de Entidades, Refinanciado, Cuenta Individual, Nombrado, Tipo Beneficio, Sexo e Ingreso Liquido.

$$P = \frac{1}{1 + e^{-X^T B}}$$

$X^T B$ = RANGOEDAD(1)*0.281 +RANGOEDAD(2)*0.221 + Nro_DEPENDIENTES(1)*0.110 + Nro_DEPENDIENTES(2)*0.056 + Nro_DEPENDIENTES(3)*(-0.144) + CALIF_EXT_N(1)*(-4.034) + CALIF_EXT_N(2)*(-4.887) + CALIF_EXT_N(3)*(-5.701)

+NUM_ENTIDADES_N(1)*(0.241) + NUM_ENTIDADES_N(2)*(0. 317) +
NUM_ENTIDADES_N(3)*(0.033) + REFINANCIADO(1)*(-0.954) +
CTA_INDIVIDUAL(1)*(0.178) + CTA_INDIVIDUAL(2)*(0.288) +
CTA_INDIVIDUAL(3)*( -0.203) + CTA_INDIVIDUAL(4)*( -0.314) +
CTA_INDIVIDUAL(5)*( 0.202) + TIPO_BENEF(1)*( -0.428) + SEXO(1)*( -0.336) +
INGRESO_LIQUIDO(1)*( 0.188)+ INGRESO_LIQUIDO(2)*(0 .741) +
INGRESO_LIQUIDO(3)*( 0.875)+ INGRESO_LIQUIDO(4)* 1.317 +
INGRESO_LIQUIDO(5)* 1.423 + INGRESO_LIQUIDO(6)* 1.340 +
INGRESO_LIQUIDO(7)* 1.005 + INGRESO_LIQUIDO(8)* 0.316 + 6.952.

La interpretación del modelo es la siguiente de acuerdo a su odds-ratio o razón de probabilidades se tiene:

En la variable Explicativa sexo se tiene un odds-ratio de 0.850 esto quiere decir que manteniendo constante los demás factores el que sea mujer SEXO(1) la probabilidad de ser un buen pagador disminuye en un factor de 0.336.

En la tabla de clasificación se puede apreciar el poder de pronóstico del modelo resultante, para la muestra de entrenamientoclasifica el modelo de regresión logística el 83% de mal pagador y el 74% de buenos pagadores. Haciendo una efectividad total de 79% la cual es bastante aceptable para modelos de comportamiento.

Tabla N° 4..2.5
Tabla de Clasificación

| Observado | | | Pronosticado | | |
|---|---|---|---|---|---|
| | | | CALIFICACION | | |
| | | | MAL PAGADOR | BUEN PAGADOR | Porcentaje Correcto |
| Entrenamiento | CALIFICACION | MAL PAGADOR | 4479 | 896 | 83% |
| | | BUEN PAGADOR | 1399 | 3976 | 74% |
| | Porcentaje global | | | | 79% |

Perfil del cliente bueno

Con el resultado de las variables significativas ya se tiene el perfil de los clientes buenos:

- El perfil del cliente bueno lo conforman los docentes de 43 a 60 años a más, con solo un dependiente, con Calificación Normal, con deuda en ninguna EntidadFinanciera, con Cuenta Individual menor a 5399, sin Refinanciamiento, de Sexo Femenino y con un Ingreso Liquido entre [851 - 970].

Construcción de la Scorecard

Se construye a continuación la scorecard. Primero se construyen los odds ratio y se realiza de acuerdo a la metodología de los siguientes cálculos para el Score:

$$Score = Offset + Factor * \ln(Odds) = Offset + Factor * (\hat{B}_0 + \sum \hat{B} woe_{ij})$$

$$Score = Offset + Factor * \ln(odds)$$

$$Score + P_0 = Offset + Factor * \ln(2 * Odds)$$

Cuyasoluciónes:

$$Factor = \frac{P_0}{\ln(2)}$$

$$Offset = Score - Factor * \ln(Odds)$$

Si la escala del score llega a 1000 como máximo el odds es equivalente a 60/1 se reemplaza entonces con $P_0 = 60$

Factor = 86.561 y Offset =561.369

Entonces el score es igual a 86.561 +561.369*ln(odds)

## 4.3 Validación del Modelo

A continuación se presentarán las respectivas validaciones que indicarán que el modelo resultante es adecuado, Entre ellas tenemos las pruebas de Hosmer-Lemeshow y para determinar si el modelo clasifica bien tendremos tres pruebas: El índice de Gini, el test de Kolgomorov-Smirnov y el Índice de Divergencia.

Prueba de Hosmer y Lemeshow

La prueba de Hosmer y Lemeshow evalúa la bondad de ajuste del modelo, la interpretación del estadístico es que en nuestro resultado se observa un valor de Chi cuadrado de 12.441 con 8 glcon un p-value de 0.134>0.05 (nivel de significancia) lo cual indica que no existe diferencia entre los valores observados y estimados lo que indica que el modelo es significativo.

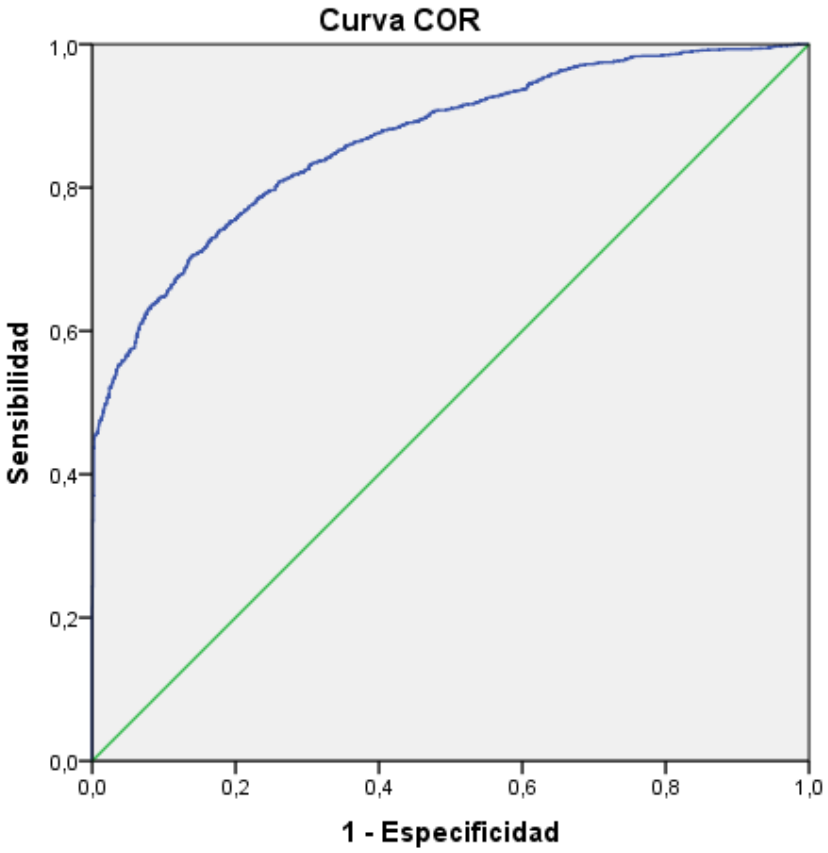
Tabla N° 4.2.6

Prueba de Hosmer y Lemeshow

| Chi-cuadrado | gl | Sig. |
|---|---|---|
| 12,411 | 8 | ,134 |

Índice Gini

El índice Gini mide la eficacia del modelo al comparar el porcentaje de clientes buenos frente al porcentaje de clientes malos para los mismos puntajes, aplicando el score a toda la población. A continuación se muestran los resultados.




Los segmentos de diagonal se generan mediante empates.

**Área bajo la curva**

Variable(s) de resultado de prueba: Probabilidad_Calculada

| Área | Error estándar[a] | Significación asintótica[b] | 95% de intervalo de confianza asintótico | |
|---|---|---|---|---|
| | | | Límite inferior | Límite superior |
| ,866 | ,006 | ,000 | ,854 | ,878 |

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

El área bajo la curva es igual a 88% lo cal indica que el modelo discrimina bien a los clientes buenos de los clientes malos.

Test de Kolgomorov- Smirnov

Una de las pruebas más utilizadas para desarrollar modelos es el índice de Kolgomorv-Smirnov que permite observar la máxima discriminación o diferencia entre clientes buenos y malos, la única diferencia es que mide las grandes separaciones en un solo punto.

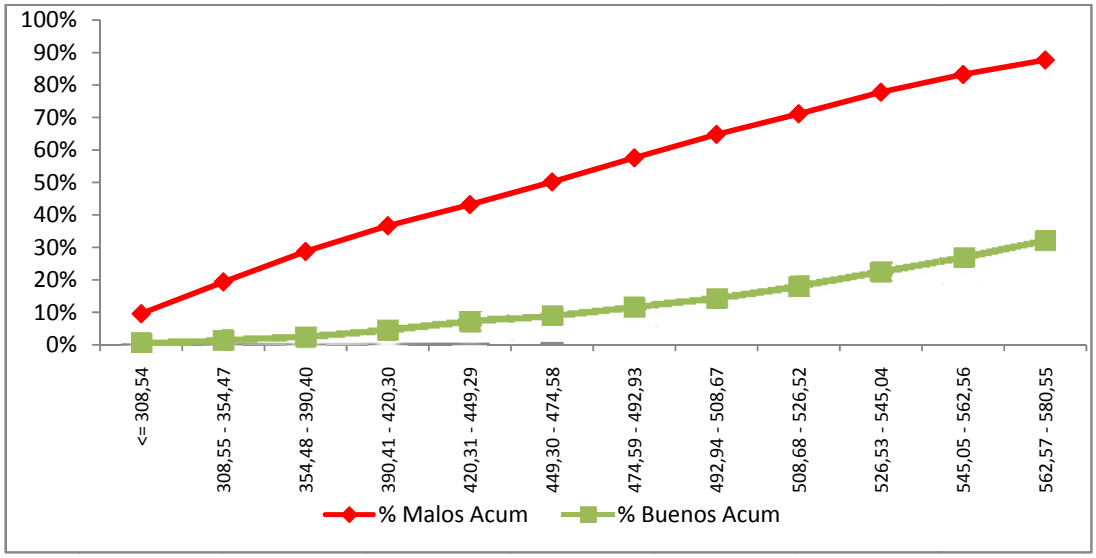
Tabla N° 4.2.7

Construcción del K-S Aplicado a la población

| | Flag_Cliente | | Total | Malos Acum | Buenos Acum | % Malos Acum | % Buenos Acum | Dif - Acumuladas |
|---|---|---|---|---|---|---|---|---|
| | Malos | Buenos | | | | | | |
| <= 308,54 | 172 | 12 | 184 | 172 | 12 | 9.6% | 0.7% | 8.9% |
| 308,55 - 354,47 | 175 | 14 | 189 | 347 | 26 | 19.4% | 1.5% | 17.9% |
| 354,48 - 390,40 | 168 | 19 | 187 | 515 | 45 | 28.7% | 2.5% | 26.2% |
| 390,41 - 420,30 | 142 | 37 | 179 | 657 | 82 | 36.7% | 4.6% | 32.1% |
| 420,31 - 449,29 | 117 | 47 | 164 | 774 | 129 | 43.2% | 7.2% | 36.0% |
| 449,30 - 474,58 | 125 | 33 | 158 | 899 | 162 | 50.2% | 9.0% | 41.1% |
| 474,59 - 492,93 | 133 | 49 | 182 | 1032 | 211 | 57.6% | 11.8% | 45.8% |
| 492,94 - 508,67 | 129 | 47 | 176 | 1161 | 258 | 64.8% | 14.4% | 50.4% |
| 508,68 - 526,52 | 114 | 66 | 180 | 1275 | 324 | 71.1% | 18.1% | 53.1% |
| 526,53 - 545,04 | 119 | 79 | 198 | 1394 | 403 | 77.8% | 22.5% | 55.3% |
| 545,05 - 562,56 | 98 | 81 | 179 | 1492 | 484 | 83.3% | 27.0% | 56.3% |
| 562,57 - 580,55 | 80 | 93 | 173 | 1572 | 577 | 87.7% | 32.2% | 55.5% |
| 580,56 - 598,93 | 80 | 88 | 168 | 1652 | 665 | 92.2% | 37.1% | 55.1% |
| 598,94 - 621,50 | 68 | 132 | 200 | 1720 | 797 | 96.0% | 44.5% | 51.5% |
| 621,51 - 653,36 | 50 | 139 | 189 | 1770 | 936 | 98.8% | 52.2% | 46.5% |
| 653,37 - 870,91 | 18 | 167 | 185 | 1788 | 1103 | 99.8% | 61.6% | 38.2% |
| 870,92 - 934,37 | 2 | 200 | 202 | 1790 | 1303 | 99.9% | 72.7% | 27.2% |
| 934,38 - 977,25 | 2 | 146 | 148 | 1792 | 1449 | 100.0% | 80.9% | 19.1% |
| 977,26 - 999,00 | 0 | 343 | 343 | 1792 | 1792 | 100.0% | 100.0% | 0.0% |

Tabla N° 4.2.8

Indice K-S Grafico de las diferencias acumuladas de Score Clientes buenos y malos




| | |
|---|---|
| **KS (SEPARACION MAXIMA)** | **56.3%** |

Construyendo el Indice K-S a toda la población de créditos tenemos que el índice resultante es de 56.3% para la muestra de validación entonces se puede determinar que este índice es bueno para discriminar a los clientes buenos de los clientes malos.

Divergencia

La última prueba de validación es la Divergencia la cual busca determinar que el grupo de clientes buenos y malos este bien separado, por lo que la diferencia de la media de sus scores deben estar bien separados. Si la divergencia es mayor a 0.95 entonces nos dirá que tenemos poblaciones estadísticamente separadas

$$Diveregencia = \frac{2 * (\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_1^2}$$

Tabla N° 4.2.9

Construcción del K-S Aplicado a la población

| | N | Media | Varianza |
|---|---|---|---|
| BUEN PAGADOR | 7,167 | 732.6635 | 44,090.37 |
| MAL PAGADOR | 7,167 | 458.2405 | 11,240.26 |

| | |
|---|---|
| Divergencia | 2.72 |

La divergencia de la muestra de validación es de 2.72 la cual es mayor a 0.95 por lo tanto las poblaciones en estudio están estadísticamente separadas.

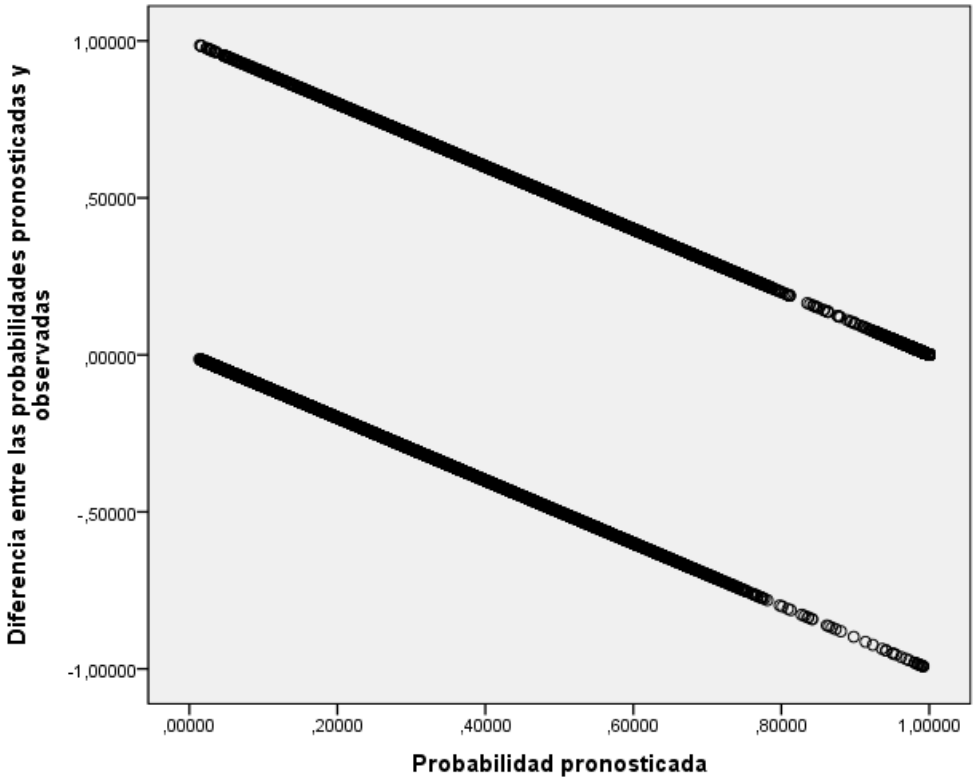
Construcción de la ScoreCard

| | | B | Exp(B) | SCORE |
|---|---|---|---|---|
| RANGOEDAD | [18 - 42] | | | |
| RANGOEDAD(1) | [43 - 60] | .281 | 1.324 | 8 |
| RANGOEDAD(2) | [61 - mas] | .221 | 1.247 | 6 |
| TENENCIA_VIVIENDA | Propia | | | |
| TENENCIA_VIVIENDA(1) | Familiar | .067 | 1.070 | 2 |
| TENENCIA_VIVIENDA(2) | Alquilada | -.158 | .854 | -5 |
| Nro_DEPENDIENTES | 0 | | | |
| Nro_DEPENDIENTES(1) | 1 | .110 | 1.117 | 3 |
| Nro_DEPENDIENTES(2) | 2 | .056 | 1.057 | 2 |
| Nro_DEPENDIENTES(3) | 3 a mas | -.144 | .866 | -4 |
| TIPO_ULT_OTOR | Consumo | | | |
| TIPO_ULT_OTOR(1) | Rapicash, Vivienda | -.236 | .789 | -7 |
| MON_ULT_OTOR | [0 - 3360] | | | |
| MON_ULT_OTOR(1) | [3361 - 4999] | -.058 | .943 | -2 |
| MON_ULT_OTOR(2) | [5000 - 6000] | -.184 | .832 | -5 |
| MON_ULT_OTOR(3) | [6001 - 18000] | -.362 | .697 | -10 |
| CANT_CRED_VIG | 0 | | | |
| CANT_CRED_VIG(1) | 1 | -3.861 | .021 | -111 |
| CANT_CRED_VIG(2) | 2 a mas | -3.807 | .022 | -110 |
| CALIF_EXT_N | NINGUNO | | | |
| CALIF_EXT_N(1) | NOR | -4.034 | .018 | -116 |
| CALIF_EXT_N(2) | CPP, DEF | -4.887 | .008 | -141 |
| CALIF_EXT_N(3) | DUD, PER | -5.701 | .003 | -165 |
| NUM_ENTIDADES_N | 0 | | | |
| NUM_ENTIDADES_N(1) | 1 | .241 | 1.272 | 7 |
| NUM_ENTIDADES_N(2) | 2 | .317 | 1.373 | 9 |
| NUM_ENTIDADES_N(3) | 3 | .033 | 1.034 | 1 |
| NUM_ENTIDADES_N(4) | 4 a mas | | | 0 |
| REFINANCIADO | NO | | | |
| REFINANCIADO(1) | SI | -.954 | .385 | -28 |
| CTA_INDIVIDUAL | 0 | | | |
| CTA_INDIVIDUAL(1) | [1 - 4392] | .178 | 1.195 | 5 |
| CTA_INDIVIDUAL(2) | [4393 - 5399] | .288 | 1.333 | 8 |
| CTA_INDIVIDUAL(3) | [5400 - 7770] | -.203 | .816 | -6 |

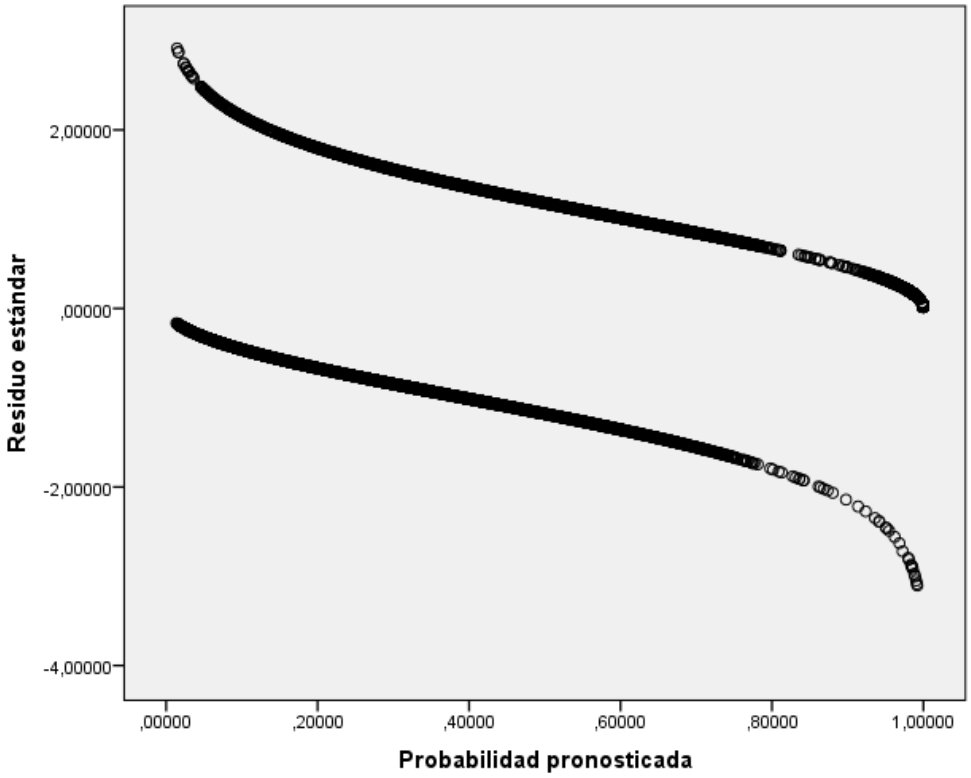
| | | | | |
|---|---|---|---|---|
| CTA_INDIVIDUAL(4) | [7771 - 9060] | -.314 | .730 | -9 |
| CTA_INDIVIDUAL(5) | [9061 - mas] | .202 | 1.224 | 6 |
| TIPO_BENEF | NINGUNO | | | |
| TIPO_BENEF(1) | INVALIDEZ, RETIRO | -.428 | .652 | -12 |
| SEXO | F | | | |
| SEXO(1) | M | -.336 | .715 | -10 |
| INGRESO_LIQUIDO | [0 - 363] | | | |
| INGRESO_LIQUIDO(1) | [364 - 489] | .188 | 1.207 | 5 |
| INGRESO_LIQUIDO(2) | [490 - 603] | .741 | 2.098 | 21 |
| INGRESO_LIQUIDO(3) | [604 - 731] | .875 | 2.398 | 25 |
| INGRESO_LIQUIDO(4) | [732 - 850] | 1.317 | 3.730 | 38 |
| INGRESO_LIQUIDO(5) | [851 - 970] | 1.423 | 4.151 | 41 |
| INGRESO_LIQUIDO(6) | [971 - 1355] | 1.340 | 3.819 | 39 |
| INGRESO_LIQUIDO(7) | [1356 - 1563] | 1.005 | 2.732 | 29 |
| INGRESO_LIQUIDO(8) | [1564 - mas] | .316 | 1.372 | 9 |

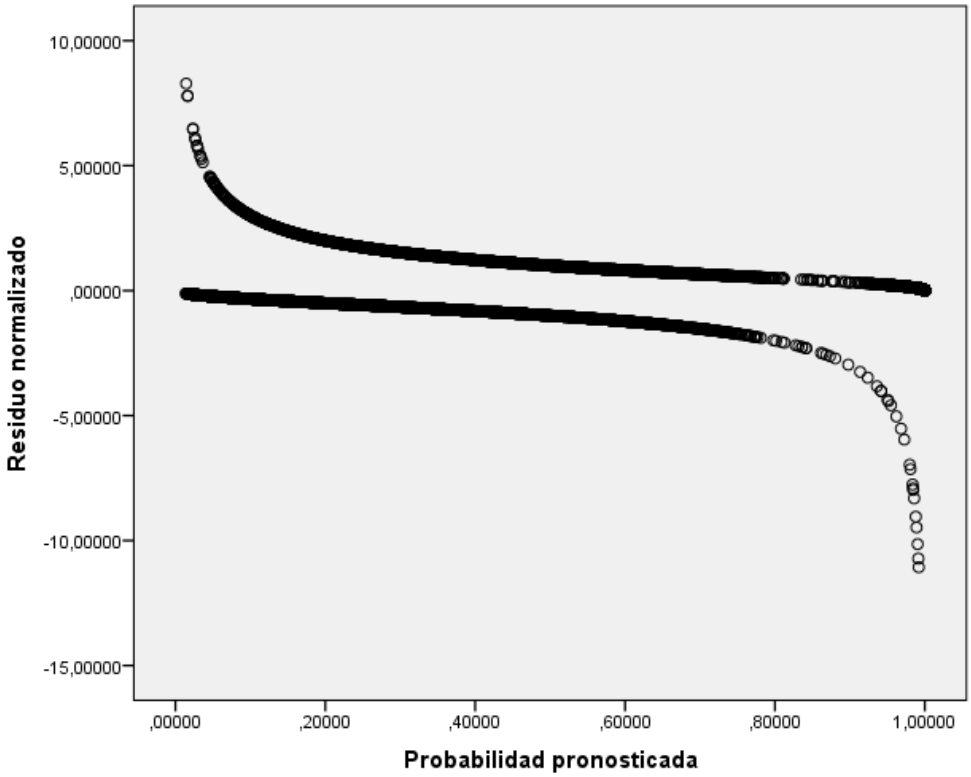
Análisis de Residuos

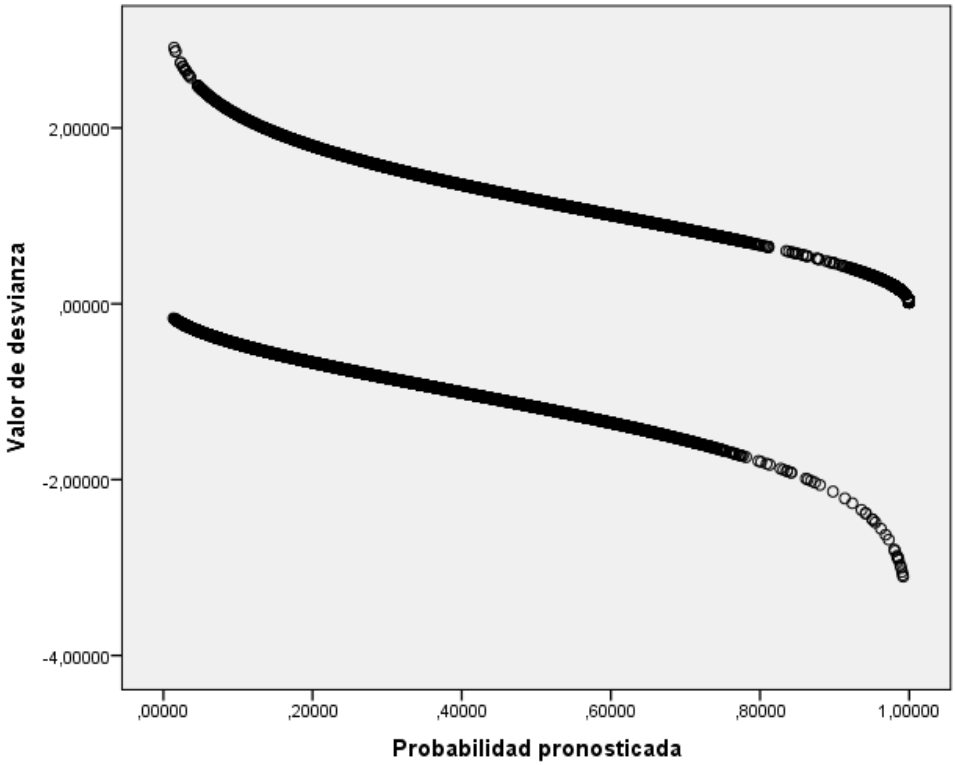
Una buena manera de ver los residuos es graficarlos contra las probabilidades predichas o simplemente con los números de casos. Dado que el tamaño de la muestra es grande se puede graficar frente a la probabilidad logística estimada. El primer grafico grafica los residuos ordinarios versus la probabilidad estimada. Los diferentes gráficos en la siguiente Figura indican dos tendencias de residuos decrecientes con pendiente -1, esto resulta del hecho de que los residuos toman solo uno de dos valores en un punto Xi.



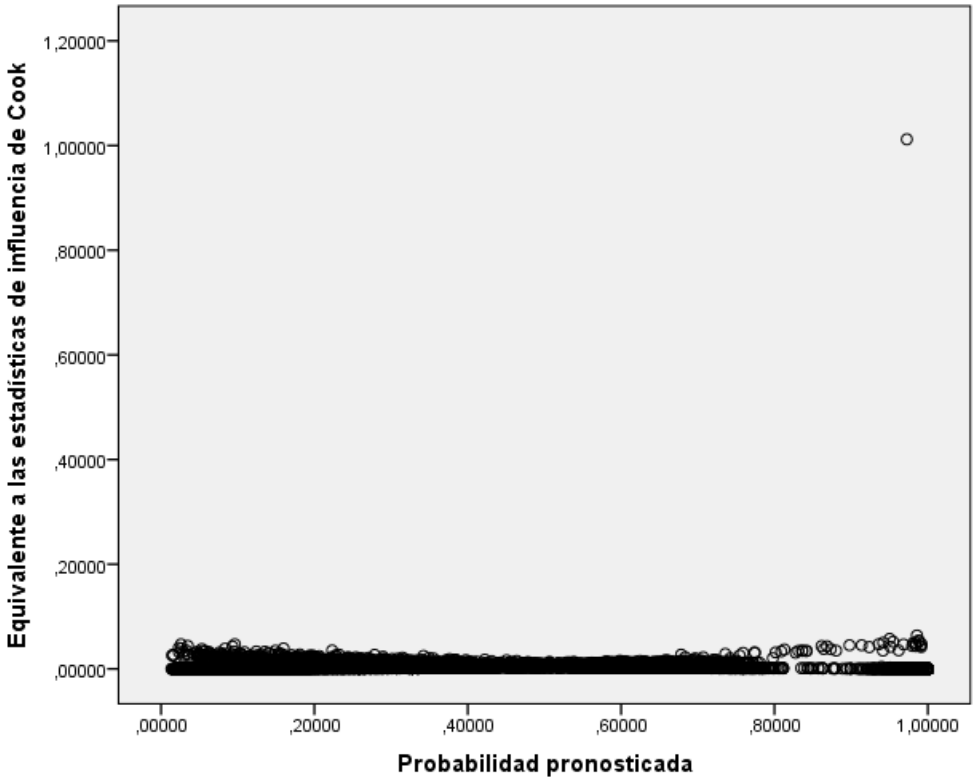

Los demás gráficos se observan un conjunto de residuales con valores menores a -2 y mayores a +2 lo cual se interpreta como valores discordantes y merecen una inspección adicional porque los valores estandarizados fuera de ese rango son valores atípicos potenciales.


Residuo estándar
2,00000
,00000
-2,00000
-4,00000
,00000
,20000
,40000
,60000
,80000
1,00000
Probabilidad pronosticada


Residuo normalizado
10,00000
5,00000
,00000
-5,00000
-10,00000
-15,00000
,00000
,20000
,40000
,60000
,80000
1,00000
Probabilidad pronosticada




Del mismo modo, de la gráfica de las distancias de Cook se observa que no existen valores discordantes o influyentes.


Equivalente a las estadísticas de influencia de Cook
1,20000
1,00000
,80000
,60000
,40000
,20000
,00000
,00000
,20000
,40000
,60000
,80000
1,00000
Probabilidad pronosticada

## 4.4 Discusión de los resultados

Discusión de los resultados

- El perfil del cliente bueno lo conforman los docentes de 43 a 60 años a más, con solo un dependiente, con Calificación Normal, con deuda en ninguna Entidad Financiera, con Cuenta Individual menor a 5399, sin Refinanciamiento, de Sexo Femenino y con un Ingreso Liquido entre [851 - 970].
- Los estadísticos de validación del modelo son importantes, es por ello que se evaluaron tres estadísticos de validación: El índice de Gini resulto ser de 89%, el test de K-S resulto ser de 56.3%y el de Divergencia de 4.92 en todos los casos el modelo de scoring paso las validaciones que concluye que el modelo desarrollado es bueno, por lo tanto clasifica muy bien los clientes buenos de los cliente malos.

Conclusiones

- El modelo construido se inició con 25 variables de las cuales solo 10 variables resultaron significativas. El modelo final se presenta a continuación.

$$P = \frac{1}{1 + e^{-X^T B}}$$

$X^T B$ = RANGOEDAD(1)*0.281 +RANGOEDAD(2)*0.221+
Nro_DEPENDIENTES(1)*0.110+ Nro_DEPENDIENTES(2)*0.056 +
Nro_DEPENDIENTES(3)*(-0.144) +CALIF_EXT_N(1)*(-4.034)+ CALIF_EXT_N(2)*(-4.887)
+ CALIF_EXT_N(3)*(-5.701)+NUM_ENTIDADES_N(1)*(0.241)+
NUM_ENTIDADES_N(2)*(0. 317)+ NUM_ENTIDADES_N(3)*(0.033) +
REFINANCIADO(1)*(-
0.954)+CTA_INDIVIDUAL(1)*(0.178)+CTA_INDIVIDUAL(2)*(0.288)+
CTA_INDIVIDUAL(3)*( -0.203)+ CTA_INDIVIDUAL(4)*( -0.314)+
CTA_INDIVIDUAL(5)*( 0.202)+ TIPO_BENEF(1)*( -0.428)+ SEXO(1)*( -0.336)+
INGRESO_LIQUIDO(1)*( 0.188)+ INGRESO_LIQUIDO(2)*(0 .741)+
INGRESO_LIQUIDO(3)*( 0.875)+ INGRESO_LIQUIDO(4)* 1.317+
INGRESO_LIQUIDO(5)* 1.423+ INGRESO_LIQUIDO(6)* 1.340+ INGRESO_LIQUIDO(7)*
1.005+ INGRESO_LIQUIDO(8)* 0.316+6.952.

En la tabla de clasificación se puede apreciar el poder de pronóstico del modelo resultante, para la muestra de entrenamiento clasifica el modelo de regresión logística el 83% de mal pagador y el 74% de buenos pagadores. Haciendo una

efectividad total de 79% la cual es bastante aceptable para modelos de comportamiento.

- Las variables significativas son: Rango edad, Numero de dependientes, Calificación Externa, Número de Entidades, Refinanciado, Cuenta Individual, Nombrado, Tipo de Beneficio, Sexo e Ingreso Liquido.

Recomendaciones

- Se sugiere desarrollar modelos de scoring en otros departamentos del país debido a que cada región tiene comportamientos de pagos diferenciados.
- Se sugiere tener precaución a la hora de recolectar la información para obtener resultados fiables, ellos parte de la recopilación de información al momento de otorgar el crédito.
- Se sugiere validar el modelo por lo menos cada dos años para así evitar que el modelo se deteriore y que la clasificación pierda efectividad.

## Bibliografía

### Revistas Científicas y Trabajos de Tesis


(1) D. García Pérez de Lema, A. Arqués Pérez y A. Calvo-Flores Segura. (Enero - Marzo 1,995). Un modelo Discriminante para evaluar el riesgo bancario en los créditos a empresas. Revista Española de Financiación y Contabilidad, Vol. XXIV, pag. 175 - 200.

(2) Geraldine Judith Vigo Chacón (2,010). *Método de Clasificación para evaluar el Riesgo crediticio: una comparación*. (Tesis para optar el título profesional de Licenciada en Estadística). Facultad de Ciencias Matemáticas. Universidad Nacional Mayor de San Marcos.

(3) M. Jesús Mures Quintana, Ana García Gallego, M. Eva Vallejo Pascual. (2,005). Aplicación del Análisis Discriminante y Regresión Logística en el estudio de la morosidad en las entidades financieras. Comparación de resultados. Revista científica de la Facultad de Ciencias Económicas y Empresariales de la Universidad de León, Vol. 1, p. 175 - 199.

(4) GutierrezGirault, Matías Alfredo (2007). Modelos de creditscoring: qué, cómo, cúando y para qué

(5) Gujarati, D. N . (2004) Econometría. McGraw-Hill Interamericana, 4ta edición, México

(6) Lyn C. Thomas, David B. Edelman y Jonathan N. Crook. "Credit Scoring and ITS Applictions". SIAM 2002 p 263

(7) Montgomery, D. (2002). Introducción al Analisis de Regresión Lineal. Compañía Editorial Continental, 3er edición, México

(8) Peña Daniel (2002) Analisis de datos multivariados . McGraw Hill, Madrid

(9) Siddiqi,Naeem (2006). Credit Risk Scorecards:developing and implementing intelligent credit scoring. John Wiley&Sons , New Jersey.

(10) Simbaqueba Lilian (2004) ¿Que es el scoring? Una vision práctica de la gestión del riesgo de crédito. Instituto del riesgo financiero, Bogotá.


### Reglamento Legal


(11) Perú. Superintendencia de Banca, Seguros y AFP's (2,008). Resolución SBS N° 11356-2008.

(12) Perú. Ministerio de Economía y Finanzas (2,014). Ley N° 30114 "Ley de Presupuesto del Sector Público"

(13) Perú. Ministerio de Economía y Finanzas (2,014). DS-N° 010-2014

(14) Perú. Superintendencia de Banca, Seguros y AFP's (2,008). Resolución S.B.S. N° 37 -2008: Reglamento de la Gestión Integral de Riesgos.

| **TÍTULO**:Aplicación de la regresión logística para la clasificación de clientes con crédito consumo del Sector Magisterial. | | | | |
|---|---|---|---|---|
| **PROBLEMA** | **OBJETIVOS** | **HIPÓTESIS** | **JUSTIFICACIÓN** | **METODOLOGÍA** |
| Problema General<br>No existe un modelo estadístico que permita clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria según su probabilidad de impago. | Objetivo general<br>Elaborar un modelo de regresión logística que permita clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria, según su probabilidad de impago. | Hipótesis principal<br>El modelo de regresión logística permita clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria, según su probabilidad de impago, con una proporción mayor a 50%. | La presente investigación resulta importante debido a que será realizada en beneficio de la Entidad Financiera. | La presente investigación es de tipo cuantitativa, descriptiva y aplicada, prospectiva y de corte transversal, que permitirá estimar el modelo scoring el cual pronostique la probabilidad de incumplimiento de impago. |
| Problema específico N° 1<br>Hasta el momento no se ha utilizado un modelo estadístico para determinar que variables son significativas para clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria según su probabilidad de impago. | Objetivo específico N° 1<br>Determinar que variables son significativas en el modelo de regresión logística para clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria según su probabilidad de impago. | Hipótesis secundaria N° 1:<br>Las variables tipo de docente, régimen de pensión, edad, estado civil, ingresos económicos y tipo de servidor son las más significativas en el modelo de regresión logística para clasificar a los clientes que solicitan crédito tipo consumo en una entidad bancaria, según su probabilidad de impago. | | Población:<br>En esta investigación, la población objetivo está comprendida por la base de datos de la cartera de créditos de la Entidad Financiera que contiene la información comportamiento de pago histórico de los clientes que solicitaron crédito durante los años 2013-2014. |

Anexos

### 4.4.1 Análisis Univariado

**Edad:**

Tabla N° 4.1
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE EDAD

| RANGO EDAD | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| [18 - 42] | 1,302 | 9.1% | 9.1% | 9.1% |
| [43 - 60] | 8,641 | 60.3% | 60.3% | 69.4% |
| [61 - mas] | 4,391 | 30.6% | 30.6% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Tipo de Docente:**

Tabla N° 4.2
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE TIPO DOCENTE

| TIPO DOCENTE | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| CESADO | 3,785 | 26.4% | 26.4% | 26.4% |
| DOCENTE | 10,549 | 73.6% | 73.6% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Estado Civil:**

Tabla N° 4.3
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE ESTADO CIVIL

| ESTADO_CIVIL | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| SOLTERO, DIVORCIADO, VIUDO | 7,421 | 51.8% | 51.8% | 51.8% |
| CASADO | 6,913 | 48.2% | 48.2% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Régimen de Pensión:**

Tabla N° 4.4

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE REGIMEN DE PENSION

| REG_PENSION | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| AFP | 6,537 | 45.6% | 45.6% | 45.6% |
| D.L. 19990 | 3,909 | 27.3% | 27.3% | 72.9% |
| D.L. 20530 | 3,888 | 27.1% | 27.1% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Tenencia de Vivienda:**

Tabla N° 4.5

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE TENENCIA DE VIVIENDA

| TENENCIA_VIVIENDA | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| Propia | 9,021 | 62.9% | 62.9% | 62.9% |
| Familiar | 3,940 | 27.5% | 27.5% | 90.4% |
| Alquilada | 1,373 | 9.6% | 9.6% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Número de Dependientes:**

Tabla N° 4.6

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE NUMERO DEPENDIENTES

| Nro_DEPENDIENTES | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| 0 | 7,654 | 53.4% | 53.4% | 53.4% |
| 1 | 2,468 | 17.2% | 17.2% | 70.6% |
| 2 | 2,655 | 18.5% | 18.5% | 89.1% |
| 3 a mas | 1,557 | 10.9% | 10.9% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Tipo Último Crédito:**

Tabla N° 4.7

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE TIPO ÚLTIMO CREDITO

| TIPO_ULT_OTOR | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| Consumo | 13,818 | 96.4% | 96.4% | 96.4% |
| Rapicash, Vivienda | 516 | 3.6% | 3.6% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Monto Último Crédito:**

Tabla N° 4.8

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE MONTO ÚLTIMO CREDITO

| MON_ULT_OTOR | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| [0 - 3360] | 2,293 | 16.0% | 16.0% | 16.0% |
| [3361 - 4999] | 2,617 | 18.3% | 18.3% | 34.3% |
| [5000 - 6000] | 7,489 | 52.2% | 52.2% | 86.5% |
| [6001 - 18000] | 1,935 | 13.5% | 13.5% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Cantidad de Créditos Vigentes:**

Tabla N° 4.9

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE CANTIDAD CREDITOS VIGENTES

| CANT_CRED_VIG | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| 0 | 3,380 | 23.6% | 23.6% | 23.6% |
| 1 | 10,407 | 72.6% | 72.6% | 96.2% |
| 2 a mas | 547 | 3.8% | 3.8% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Calificación Externa:**

Tabla N° 4.11

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE CALIFICACIÓN EXTERNA

| CALIF_EXT_N | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| NINGUNO | 1,272 | 8.9% | 8.9% | 8.9% |
| NOR | 10,145 | 70.8% | 70.8% | 79.6% |
| CPP, DEF | 954 | 6.7% | 6.7% | 86.3% |
| DUD, PER | 1,963 | 13.7% | 13.7% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Número de Entidades:**

Tabla N° 4.12

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE NUMERO DE ENTIDADES

| NUM_ENTIDADES_N | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| 0 | 1,272 | 8.9% | 8.9% | 8.9% |
| 1 | 2,785 | 19.4% | 19.4% | 28.3% |
| 2 | 3,324 | 23.2% | 23.2% | 51.5% |
| 3 | 2,812 | 19.6% | 19.6% | 71.1% |
| 4 a mas | 4,141 | 28.9% | 28.9% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Refinanciado:**

Tabla N° 4.13

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE REFINANCIADO

| REFINANCIADO | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| NO | 11,705 | 81.7% | 81.7% | 81.7% |
| SI | 2,629 | 18.3% | 18.3% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Cuenta Individual:**

Tabla N° 4.14

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE CUENTA INDIVIDUAL

| CTA_INDIVIDUAL | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|
| 0 | 3,783 | 26.4% | 26.4% | 26.4% |
| [1 - 4392] | 1,976 | 13.8% | 13.8% | 40.2% |
| [4393 - 5399] | 1,344 | 9.4% | 9.4% | 49.6% |
| [5400 - 7770] | 1,416 | 9.9% | 9.9% | 59.4% |
| [7771 - 9060] | 3,104 | 21.7% | 21.7% | 81.1% |
| [9061 - mas] | 2,711 | 18.9% | 18.9% | 100.0% |
| Total | 14,334 | 100.0% | 100.0% | |

**Nombrado:**

Tabla N° 4.15

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE NOMBRADO

| NOMBRADO | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | NO | 887 | 6.2% | 6.2% | 6.2% |
| | SI | 13,447 | 93.8% | 93.8% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

**Tipo de Beneficio:**

Tabla N° 4.16

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE TIPO BENEFICIO

| TIPO_BENEF | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | NINGUNO | 13,642 | 95.2% | 95.2% | 95.2% |
| | INVALIDEZ, RETIRO | 692 | 4.8% | 4.8% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

**Género:**

Tabla N° 4.17
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE GÉNERO

| GENERO | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | F | 9,301 | 64.9% | 64.9% | 64.9% |
| | M | 5,033 | 35.1% | 35.1% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

**Ingreso Líquido:**

Tabla N° 4.18
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE INGRESO LÍQUIDO

| INGRESO_LIQUIDO | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | [0 - 363] | 2634 | 18% | 18% | 18% |
| | [364 - 489] | 2182 | 15% | 15% | 34% |
| | [490 - 603] | 1575 | 11% | 11% | 45% |
| | [604 - 731] | 1459 | 10% | 10% | 55% |
| | [732 - 850] | 1122 | 8% | 8% | 63% |
| | [851 - 970] | 1105 | 8% | 8% | 70% |
| | [971 - 1355] | 2060 | 14% | 14% | 85% |
| | [1356 - 1563] | 925 | 6% | 6% | 91% |
| | [1564 - mas] | 1272 | 9% | 9% | 100% |
| | Total | 14334 | 100% | 100% | |

**Variable Respuesta Calificación:**

Tabla N° 4.20
DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE RESPUESTA

| CALIFICACION | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | MAL PAGADOR | 7,167 | 50.0% | 50.0% | 50.0% |
| | BUEN PAGADOR | 7,167 | 50.0% | 50.0% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

**Tipo de Muestra:**

Tabla N° 4.21

DISTRIBUCIÓN DE FRECUENCIA DE LA VARIABLE RESPUESTA

| TIPO MUESTRA | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válido | VERIFICACION | 3,584 | 25.0% | 25.0% | 25.0% |
| | ENTRENAMIENTO | 10,750 | 75.0% | 75.0% | 100.0% |
| | Total | 14,334 | 100.0% | 100.0% | |

### 4.4.2 Análisis Bivariado

**Edad:**

Tabla N° 4.22

EDAD VS VARIABLE RESPUESTA

**RANGO EDAD*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| RANGO EDAD | [18 - 42] | 694 | 608 | 1302 |
| | [43 - 60] | 4260 | 4381 | 8641 |
| | [61 - mas] | 2213 | 2178 | 4391 |
| Total | | 7167 | 7167 | 14334 |

**Tipo de docente:**

Tabla N° 4.23

TIPO DE DOCENTE VS VARIABLE RESPUESTA

**TIPO DOCENTE*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| TIPO DOCENTE | CESADO | 1963 | 1822 | 3785 |
| | DOCENTE | 5204 | 5345 | 10549 |
| Total | | 7167 | 7167 | 14334 |

**Estado Civil:**

Tabla N° 4.24

ESTADO CIVIL VS VARIABLE RESPUESTA

**ESTADO_CIVIL*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| ESTADO_CIVIL | SOLTERO, DIVORCIADO, VIUDO | 3697 | 3724 | 7421 |
| | CASADO | 3470 | 3443 | 6913 |
| Total | | 7167 | 7167 | 14334 |

**Régimen de Pensión:**

Tabla N° 4.25

REGIMEN DE PENSION VS VARIABLE RESPUESTA

**REG_PENSION*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| REG_PENSION | AFP | 3360 | 3177 | 6537 |
| | D.L. 19990 | 1780 | 2129 | 3909 |
| | D.L. 20530 | 2027 | 1861 | 3888 |
| Total | | 7167 | 7167 | 14334 |

**Tenencia de Vivienda:**

Tabla N° 4.26

TENENCIA DE VIVIENDA VS VARIABLE RESPUESTA

**TENENCIA_VIVIENDA*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| TENENCIA_VIVIENDA | Propia | 4375 | 4646 | 9021 |
| | Familiar | 2047 | 1893 | 3940 |

| | | | | |
|---|---|---|---|---|
| | Alquilada | 745 | 628 | 1373 |
| Total | | 7167 | 7167 | 14334 |

**N° Dependientes:**

Tabla N° 4.27

NUMERO DE DEPENDIENTES VS VARIABLE RESPUESTA

**Nro_DEPENDIENTES*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| Nro_DEPENDIENTES | 0 | 3912 | 3742 | 7654 |
| | 1 | 1138 | 1330 | 2468 |
| | 2 | 1295 | 1360 | 2655 |
| | 3 a mas | 822 | 735 | 1557 |
| Total | | 7167 | 7167 | 14334 |

**Tipo de Último Crédito:**

Tabla N° 4.28

TIPO ÚLTIMO CREDITO VS VARIABLE RESPUESTA

**TIPO_ULT_OTOR*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| TIPO_ULT_OTOR | Consumo | 6874 | 6944 | 13818 |
| | Rapicash, Vivienda | 293 | 223 | 516 |
| Total | | 7167 | 7167 | 14334 |

**Monto de Último Crédito:**

Tabla N° 4.29

MONTO ÚLTIMO CREDITO VS VARIABLE RESPUESTA

**MON_ULT_OTOR*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| MON_ULT_OTOR | [0 - 3360] | 718 | 1575 | 2293 |
| | [3361 - 4999] | 1044 | 1573 | 2617 |
| | [5000 - 6000] | 4074 | 3415 | 7489 |
| | [6001 - 18000] | 1331 | 604 | 1935 |
| Total | | 7167 | 7167 | 14334 |

**Cantidad de Créditos Vigentes:**

Tabla N° 4.30

CANTIDAD CREDITOS VIGENTES VS VARIABLE RESPUESTA

**CANT_CRED_VIG*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| CANT_CRED_VIG | 0 | 56 | 3324 | 3380 |
| | 1 | 6778 | 3629 | 10407 |
| | 2 a mas | 333 | 214 | 547 |
| Total | | 7167 | 7167 | 14334 |

**Calificación Externa:**

Tabla N° 4.32

CALIFICACION EXTERNA VS VARIABLE RESPUESTA

**CALIF_EXT_N*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| CALIF_EXT_N | NINGUNO | 1 | 1271 | 1272 |
| | NOR | 4746 | 5399 | 10145 |
| | CPP, DEF | 738 | 216 | 954 |

| | | | | |
|---|---|---|---|---|
| | DUD, PER | 1682 | 281 | 1963 |
| Total | | 7167 | 7167 | 14334 |

**Numero Entidades:**

Tabla N° 4.33

NUMERO DE ENTIDADES VS VARIABLE RESPUESTA

**NUM_ENTIDADES_N*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| NUM_ENTIDADES_N | 0 | 1 | 1271 | 1272 |
| | 1 | 986 | 1799 | 2785 |
| | 2 | 1630 | 1694 | 3324 |
| | 3 | 1752 | 1060 | 2812 |
| | 4 a mas | 2798 | 1343 | 4141 |
| Total | | 7167 | 7167 | 14334 |

**Refinanciado:**

Tabla N° 4.34

REFINANCIADO VS VARIABLE RESPUESTA

**REFINANCIADO*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| REFINANCIADO | NO | 4899 | 6806 | 11705 |
| | SI | 2268 | 361 | 2629 |
| Total | | 7167 | 7167 | 14334 |

**Cuenta Individual:**

Tabla N° 4.35

CUENTA INDIVIDUAL VS VARIABLE RESPUESTA

**CTA_INDIVIDUAL*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| CTA_INDIVIDUAL | 0 | 1961 | 1822 | 3783 |
| | [1 - 4392] | 969 | 1007 | 1976 |
| | [4393 - 5399] | 588 | 756 | 1344 |
| | [5400 - 7770] | 719 | 697 | 1416 |
| | [7771 - 9060] | 1682 | 1422 | 3104 |
| | [9061 - mas] | 1248 | 1463 | 2711 |
| Total | | 7167 | 7167 | 14334 |

**Nombrado:**

Tabla N° 4.36

NOMBRADO VS VARIABLE RESPUESTA

**NOMBRADO*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| NOMBRADO | NO | 397 | 490 | 887 |
| | SI | 6770 | 6677 | 13447 |
| Total | | 7167 | 7167 | 14334 |

**Tipo Beneficio:**

Tabla N° 4.37

TIPO BENEFICIO VS VARIABLE RESPUESTA

**TIPO_BENEF*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| TIPO_BENEF | NINGUNO | 6782 | 6860 | 13642 |
| | INVALIDEZ, RETIRO | 385 | 307 | 692 |
| Total | | 7167 | 7167 | 14334 |

**Género:**

Tabla N° 4.38

GENERO VS VARIABLE RESPUESTA

**SEXO*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| SEXO | F | 4375 | 4926 | 9301 |
| | M | 2792 | 2241 | 5033 |
| Total | | 7167 | 7167 | 14334 |

**Ingreso Líquido:**

Tabla N° 4.39

INGRESO LIQUIDO VS VARIABLE RESPUESTA

**INGRESO_LIQUIDO*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| INGRESO_LIQUIDO | [0 - 363] | 2068 | 566 | 2634 |
| | [364 - 489] | 1573 | 609 | 2182 |
| | [490 - 603] | 951 | 624 | 1575 |
| | [604 - 731] | 727 | 732 | 1459 |
| | [732 - 850] | 358 | 764 | 1122 |
| | [851 - 970] | 304 | 801 | 1105 |
| | [971 - 1355] | 499 | 1561 | 2060 |
| | [1356 - 1563] | 154 | 771 | 925 |
| | [1564 - mas] | 533 | 739 | 1272 |
| Total | | 7167 | 7167 | 14334 |

**Tipo Muestra:**

Tabla N° 4.40

TIPO MUESTRA VS VARIABLE RESPUESTA

**TIPO MUESTRA*CALIFICACION tabulación cruzada**

Recuento

| | | CALIFICACION | | |
|---|---|---|---|---|
| | | MAL PAGADOR | BUEN PAGADOR | Total |
| TIPO MUESTRA | VERIFICACION | 1792 | 1792 | 3584 |
| | ENTRENAMIENTO | 5375 | 5375 | 10750 |
| Total | | 7167 | 7167 | 14334 |