UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS

POSTGRADO EN MODELOS ALEATORIOS

DESARROLLO DE UN *SCORE* DE CRÉDITO UTILIZANDO ANÁLISIS CARACTERÍSTICO Y ÁRBOLES DE CLASIFICACIÓN PARA INFERIR LA PROBABILIDAD DE INCUMPLIMIENTO EN LOS CLIENTES NUEVOS DE UNA ENTIDAD BANCARIA

TRABAJO DE GRADO DE MAESTRÍA

A ser desarrollado por la:
Licda. Emily Carolina Piñero Arreaza
Firma: _____

Tutor: Dr. José Benito Hernández
Lugar de trabajo: UCV
Firma: _____

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
COMISIÓN DE ESTUDIOS DE POSTGRADO

Comisión de Estudios de Postgrado

VEREDICTO

Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Ciencias de la Universidad Central de Venezuela, para examinar el **Trabajo de Grado** presentado por: Emily Carolina Piñero Arreaza, **Cédula de identidad** 18.810.887, bajo el título "**DESARROLLO DE UN SCORE DE CRÉDITO UTILIZANDO ANÁLISIS CARACTERÍSTICO Y ÁRBOLES DE CLASIFICACIÓN PARA INFERIR LA PROBABILIDAD DE INCUMPLIMIENTO EN LOS CLIENTES NUEVOS DE UNA ENTIDAD BANCARIA**", a fin de cumplir con el requisito legal para optar al grado académico de **MAGÍSTER SCIENTIARUM, MENCIÓN MODELOS ALEATORIOS**, dejan constancia de lo siguiente:

1.- Leído como fue dicho trabajo por cada uno de los miembros del jurado, se fijó el día 17 de enero de 2019 a las 3:00 pm., para que el autorlo defendiera en forma pública, lo que éste hizo en el **Postgrado de Matemática** mediante un resumen oral de su contenido, luego de lo cual respondió satisfactoriamente a las preguntas que le fueron formuladas por el jurado, todo ello conforme con lo dispuesto en el Reglamento de Estudios de Postgrado.

2.- Finalizada la defensa del trabajo, el jurado decidió **aprobarlo** por considerar, sin hacerse solidario con la ideas expuestas por el autor, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado.

Para dar este veredicto, el jurado estimó que el trabajo proporciona información útil para el estudio de score de créditos, estableciendo metodologías alternativas al estudio clásico de los mismos, lo que resulta un aporte a la aplicación de la estadística para el manejo de grandes volúmenes de datos.

En fe de lo cual se levanta la presente ACTA, a los 17 días del mes de enero del año 2019, conforme a lo dispuesto en el Reglamento de Estudios de Postgrado, actuó como Coordinadordel jurado el **tutor Dr. José Benito Hernández**.

Dra. Mairene Colina
C.I.12.761.954
Institución UCV

Dr. Ricardo Ríos
C.I. 3.949.476
Institución UCV

Dr. José Benito Hernández
C.I. 10.867.203
Institución UCV
Tutor

RESUMEN

Este proyecto consistió en elaborar y analizar modelos de score basados en datos reales en formato texto de una entidad bancaria argentina. Se utilizó en primera instancia el análisis característico y posteriormente árboles de clasificación, ambas como metodología para la selección de variables. Estos modelos se realizaron en el software libre R, el cual es una herramienta estadística que permite conceptualizar, analizar y optimizar modelos de regresión logística. Dicho estudio se realizó con la necesidad de implementar y comparar estas metodologías lo que proporciona la confianza necesaria para hallar el mejor conjunto de variables para una futura implementación y validación de estos modelos con datos actuales del cliente.

AGRADECIMIENTOS

Quisiera agradecer principalmente a Dios, por acompañarme y guiar mi camino para así lograr cumplir una nueva meta en mi vida. A mis padres, Celsa Arreaza y Emilio Piñero por su incondicional apoyo en todo lo que me he propuesto y por todas sus enseñanzas y consejos que han hecho lo que soy hoy en día. A mi hermana Dilia Patricia Piñero Arreaza, que más que una hermana es una amiga que siempre estará para mí cuando lo necesite.

A la Universidad Central de Venezuela, por abrirme sus puertas y permitir ser una futura egresada de tan prestigiosa casa de estudio. A sus profesores, que a pesar de momentos díficiles, se mantienen haciendo lo que realmente aman. A mi tutor, el profesor José Benito Hernández que gracias a su dedicación a la hora de enseñar, tuve la base necesaria para hoy en día realizar este trabajo de grado, por su apoyo, muchas gracias.

A Tomás León por permitirme pertenecer a tan importarte empresa como lo es Datacrédito Experian. A mis compañeros de trabajo, Jesus Yerena, Karen Coloma, Sandra Loo y Juan Pablo Vidal por brindarme la oportunidad de llevar a cabo este proyecto y permitirme que sea mi trabajo de grado. A todos ustedes muchas gracias por los consejos y ayudas.

A mis colegas y amigos Adrian, Danny y Alisbel, infinitas gracias por la paciencia, consejos y apoyo en los momentos más complicados dentro de este proyecto, ustedes lograron que pudiera ver todo más sencillo.

A todos ustedes, mi más sincero respeto y agradecimiento.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La utilización de modelos de *credit scoring* para la evaluación del riesgo de crédito, es decir, para estimar probabilidades de *default* y ordenar a los deudores y solicitantes de financiamiento en función de su riesgo de incumplimiento, comenzó en los 70's pero se generalizó a partir de los 90's. Esto se ha debido tanto al desarrollo de mejores recursos estadísticos y computacionales, como por la creciente necesidad por parte de la industria bancaria de hacer más eficaz y eficiente la generación de financiaciones, y de tener una mejor evaluación del riesgo de su portafolio. Estos modelos generalmente se asocian a todos aquellos procedimientos que permiten extraer información útil y encontrar patrones de comportamiento de los datos.

El término *credit scoring* recoge todos los métodos estadísticos que se utilizan para determinar el riesgo asociado con un posible deudor. La importancia de estos métodos crece día por día en el sector crediticio, pues nadie quiere otorgar créditos que probablemente no serán pagados a "malos deudores", ni tampoco perder la oportunidad de obtener una ganancia al otorgar un crédito a un "buen deudor". Por lo que es importante entonces utilizar métodos que clasifiquen a los deudores de manera efectiva y que descubran patrones imperceptibles que ayuden a determinar el riesgo de un aspirante.

Estos modelos usan metodologías estadísticas para clasificar y ordenar clientes buenos y malos en función a una variable respuesta (variable dependiente), mediante la asignación de puntajes diferenciales a los clientes. Dicha asignación se basa en la exploración de ciertos hechos del pasado en busca de patrones de comportamiento que presenten una estrecha relación con el evento de estudio, para posteriormente estimar la probabilidad que este evento suceda y otorgar un puntaje a cada individuo.

Este enfoque sirvió como base para llevar a cabo dicho trabajo, en donde se estudiarán dos modelos utilizando sendas metodologías que se usarán como soporte para analizar las diferencias en relación a la selección de variables que se utilizarán posteriormente para la elaboración de modelos de *scoring*. Dicho estudio se realizará con la necesidad de implementar y comparar estas metodologías en el *software* libre R, lo que proporcionará la confianza necesaria para hallar el mejor conjunto de variables para una futura implementación y validación de estos modelos con datos actuales del cliente.

En general, para dicho proyecto se propondrá el siguiente objetivo: establecer la mejor metodología entre análisis característico y árboles de clasificación que permita generar modelos de *score* y estimar la probabilidad de incumplimiento de los clientes nuevos de una cierta entidad bancaria.

Mientras que como objetivos específicos, se tendrán:

- Estudiar y adquirir las habilidades necesarias en el manejo de ambas metodologías.
- Comparar el análisis característico y los árboles de clasificación en relación a la selección de variables que serán utilizadas dentro de los modelos de *score*.
- Utilizar la población de clientes nuevos aprobados por la entidad para elaborar el modelo con el conjunto de variables proporcionadas por la metodología utilizada en Experian Soluciones V (análisis característico).
- Utilizar la población de clientes nuevos aprobados por la entidad para elaborar un segundo modelo con el conjunto de variables que se obtuvieron a traves de los árboles de clasificación con ayuda del paquete *ctree* del *software* libre R.
- Inferir la probabilidad de incumplimiento en el pago por parte de los clientes nuevos rechazados de una entidad bancaria.
- Realizar un modelo de *score* para toda la población (aprobada + rechazada) que se va a evaluar.
- Definir la escala del *score* así como el umbral que define el comportamiento (bueno o malo) de un cliente nuevo.
- Realizar pruebas de validación para comprobar que los modelos de *score* generados posean una calibración y discriminación adecuada.

CAPÍTULO 1

DESCRIPCIÓN DE LA EMPRESA

(EXPERIAN SOLUCIONES V, S.A.)

El Buró de Crédito es una empresa privada que recibe información de cada banco y entidad financiera y la transforma en un historial que archiva con el nombre de la persona (física o moral) que lo solicitó. De esta forma va recaudando datos, que comparte a través de un reporte a las empresas que están analizando otorgar un financiamiento. Esto ayuda a conocer el comportamiento del cliente y en eso basarse para tomar una decisión**.**

Experian es la compañía líder en servicios globales de la información con funcionamiento a través de 39 países, fue formalmente fundada en 1996 cuando se fusionaron *Commercial Credit Nottingham* del Reino Unido y *TRW Information Services* de Estados Unidos. Hoy día, mantiene su sede principal en Dublín, Irlanda y extiende sus servicios y líneas de negocio a través de las distintas ramificaciones de la compañía establecidas a nivel mundial.

Especificamente, Experian Soluciones V, S.A, antiguo DataCrédito, es la Central de Información Crediticia en Venezuela que cuenta con 40 años de experiencia en el mercado de soluciones de información y se basa en proveer datos y herramientas analíticas para clientes de América Latina, de esta forma, ayuda a las empresas e instituciones financieras a manejar el riesgo crediticio, segmentar carteras de clientes y ofertas de *marketing*, prevenir la suplantación de identidad y automatizar decisiones respecto a servicios financieros que ciertas organizaciones ofrecen, así mismo, también provee información a personas acerca de su valoración crediticia frente a diversas entidades financieras. En general, sus principales funciones son:

- Suministrar soluciones integrales a los principales sectores de la economía para la toma de mejores decisiones en el ciclo de otorgamiento de crédito.
- Innovar en el conocimiento y la confiabilidad del manejo de la información a través de la administración de la base de datos más completa del país con información de identificación, localización demográfica, hábito de pago y nivel de endeudamiento de personas naturales y jurídicas.
- Ofrecer a las personas permanente acceso a su historia de crédito para que puedan conocer el estado de sus obligaciones y detectar el uso fraudulento de su nombre.

Para optimizar su labor, Experian Soluciones V, S.A cuenta con áreas y departamentos, los cuales se dividen en 4 grandes grupos: *Credit Services*, *Decision Analytics*, *Marketing Services* y *Consumer Services*.

En el año 2013, Experian Soluciones V, S.A dentro de un proceso de adaptación, creó la dirección de Modelos a la Medida dentro del departamento de *Decision Analytics*, la cual está especializada en el desarrollo de soluciones y herramientas analíticas a la medida, las cuales permiten mejorar el conocimiento y entendimiento de sus clientes con el fin de optimizar sus políticas, maximizar la rentabilidad de sus estrategias y agilizar la toma de decisiones. Esta oficina, a la cual pertenezco con el cargo de Especialista en Investigación y Desarrollo, tiene como propósito ofrecer a los clientes y a las unidades de *Spanish Latam*, productos y/o servicios innovadores y sofisticados, a través de recurso humano experto y tecnología de punta, maximizando el uso de la información de buró y propia, así mismo:

- Profundizar para obtener mejores resultados en el análisis de riesgo.
- Apoyar en todas las fases del ciclo de vida de los clientes permitiendo:
  - Ser mas asertivo en la adquisición de nuevos clientes.
  - Agilizar la gestión de cuentas y la recuperación de las mismas.
  - Realizar estrategias que incrementen la rentabilidad de las operaciones.
- Proveer y recomendar soluciones con calidad que garanticen la aplicación de criterios estadísticos y matemáticos robustos.
- Garantizar el desarrollo de las actividades de acuerdo a los tiempos establecidos con los clientes.
- Mantener una revisión permanente de las mejores prácticas, a fin de identificar oportunamente metodologías y procesos que apunten a atender nuevos segmentos de negocios.

CAPÍTULO 2

MARCO TEÓRICO

Un modelo de *score* de crédito es un modelo estadístico que mide la probabilidad de que un cliente nuevo de cierta entidad bancaria caiga en mora en un período de tiempo específico. Dichos modelos estadísticos toman en consideración variables predictivas reportadas directamente por el cliente de tipo numérico y categórico, como por ejemplo, características sociales, demográficas y económicas de cada uno de los individuos de interés; variables de buró, entiéndase como, el conjunto de variables que no fueron suministradas por el cliente sino por el contrario es información que se encuentra en la Central de Información Crediticia recaudada por las demás entidades bancarias; y todas aquellas variables que describen la relación entre el cliente y el banco, aquellas que reflejan los movimientos transaccionales y que pueden explicar la característica de interés.

Para la construcción del modelo, estas variables deben atravesar un tratamiento previo con lo cual se espera que el modelo final tenga aquel conjunto de variables que mejor expliquen el fenómeno de estudio. Para el caso específico de este trabajo de investigación, se utilizarán dos análisis con el fin de hallar el mejor conjunto de variables que se utilizarán posteriormente para la elaboración de los modelos.

2.1 ANÁLISIS CARACTERÍSTICO

Antes de iniciar el proceso de modelado, se realizará un análisis característico de la información reportada por la entidad bancaria [1], esto incluye un análisis univariado, bivariado y multivariado del conjunto inicial de variables lo cual permitirá filtrar dicho conjunto a través de criterios como variabilidad, completitud, colinealidad, entre otros, para así obtener un subconjunto de variables explicativas con gran poder predictivo y relevancia en términos del negocio.

2.1.1 ANÁLISIS UNIVARIADO

En la generación de un modelo matemático, resulta importante que las variables o características que participan en éste posean información significativa, es por ello que la primera etapa en este proceso se enfoca en el estudio y transformación de las variables. La primera fase es el análisis

univariado, donde se estudia el porcentaje de completitud que poseen, ya que el mismo informa cuánta información puede aportar al modelo dicha variable. Una variable que posea un 95% o más de valores faltantes se descarta en esta etapa del proceso por considerarse incompleta.

Una vez realizado el análisis de completitud de las variables, se estudia la heterogeneidad (variabilidad) de las variables, entendiendo por esto el análisis de la dispersión o cercanía de los valores de la misma. Para decidir si la variable aporta información en el desarrollo de los modelos, es necesario contemplar la distribución de la misma, es decir, sus percentiles.

Se considerará que existe variabilidad cuando la misma supera el 2%, es decir, si P_1 y P_{99} son iguales, existe un 2% de variabilidad, por lo tanto, la variable se descarta. Para los demás casos, la variable aporta información y es considerada en el estudio.

Por ejemplo, si la variable edad se encuentra entre un rango de 18 y 60 años, y el P_1 y el P_{99} es igual a 22, se estaría considerando que el 98% de la población estudiada tiene 22 años de edad, por lo que no sería una variable adecuada para realizar el modelo ya que se explicaría sólo a las personas con esta edad.

Por otra parte, se tratan los valores atípicos por medio del acotamiento de la variable a través de percentiles, es decir, se acotará con el P_{95} como cota superior y P_5 como cota inferior, de esta manera se previene la influencia que pueden tener estos valores en la implementación de alguno de los modelos.

Luego de revisar las variables en cada uno de los diferentes aspectos, se obtendrá el total de variables a las cuales se les realizará el análisis bivariado y se conocerá también el trato que se le dará a las variables de ahora en adelante, si es de forma continua o de forma categórica.

2.1.2 ANÁLISIS BIVARIADO

En esta fase del análisis característico, aquellas variables que aprobaron el análisis univariado, es decir, aquellas variables con la completitud y variabilidad requerida, se comparan frente a la

variable BGI (*Bad, Good, Indeterminated*) , bueno, malo o indeterminado, con la finalidad de que el modelo que contenga estas variables sea capaz de tener una buena calidad de predicción en esos segmentos de clientes que se consideran buenos y malos.

La variable BGI representará la variable respuesta o dependiente del modelo que se realizará, es una construcción sujeta a la planificación conjunta entre analista - cliente y puede obedecer a estrategias comerciales de la entidad bancaria, dicha variable se construirá de la siguiente manera:

$$BGI = \begin{cases} 1 & si\, el\, cliente\, es\, malo. \\ 0 & si\, el\, cliente\, es\, bueno. \end{cases}$$

El principal objetivo de este análisis, es particionar las variables en clases, dicha partición se hace de acuerdo a la experiencia de negocio o simplemente consiste en un particionamiento lógico en el que cada clase posee una misma proporción de valores. A continuación, se puede apreciar un ejemplo de agrupación de una variable frente al BGI.

Clases	**Total Clases**	**% Total**	**# Buenos**	**% Buenos**	**# Malos**	**% Malos**	**Tasa de Malos**	**IV**
Clase I	n_1	n_1/N	B_1	B_1/B	M_1	M_1/M	M_1/n_1	IV
Clase II	n_2	n_2/N	B_2	B_2/B	M_2	M_2/M	M_2/n_2	IV
Clase III	n_3	n_3/N	B_3	B_3/B	M_3	M_3/M	M_3/n_3	IV
Totales	N	100%	B	$\frac{B_1+B_2+B_3}{B}$	M	$\frac{M_1+M_2+M_3}{M}$	$\frac{M_1}{n_1}+\frac{M_2}{n_2}+\frac{M_3}{n_3}$	

Tabla 2.1.2.1 Ejemplo de Partición de Variable

Cálculos Estadísticos

Una vez que se tienen las particiones, cada variable debe cumplir con una serie de validaciones para aprobar su continuidad en el proceso. Inicialmente se deben calcular los siguientes estadísticos para cada una de las variables analizadas:

- **WOE (*Weight of Evidence*):** Hace referencia al peso de evidencia y cuyo propósito es medir la fuerza de predicción y separación entre buenos y malos de cada clase o categoría de una

variable, esto es, la probabilidad de que una persona en una cierta clase sea buena o mala. Se calcula como sigue:

$$WOE_i = ln(\frac{\%Buenos_i}{\%Malos_i}).$$

donde:

- %Buenos: Corresponde a la frecuencia relativa de la categoría i-ésima respecto de las otras categorías para los buenos.

$$\%Buenos_i = \frac{\#Buenos_i}{Total\,Buenos}.$$

- %Malos: Corresponde a la frecuencia relativa de la categoría i-ésima respecto de las otras categorías para los malos.

$$\%Malos_i = \frac{\#Malos_i}{Total\,Malos}.$$

El total de buenos y malos, corresponde a los totales de buenos y malos que una determinada variable arroja luego de compararse con el BGI.

- **IV (*Information Value*):** O valor de información, mide la fuerza total de separación que tiene una variable en el análisis bivariado, para su cálculo requiere el cálculo del WOE.

$$IV = \sum_{i=1}^{k} \{(\%Buenos_i - \%Malos_i) * WOE_i\}.$$

donde k es la cantidad de clases en las que una variable fue particionada.

Su umbral de evaluación es el siguiente:

IV < 0.02: No predictivo,

0.02 – 0.1: Predictibilidad débil,

0.1 – 0.3: Predictibilidad media,

0.3 – 0.5: Predictibilidad fuerte,

0.5+: Debe revisarse.

En general, en el análisis bivariado se busca que la magnitud del WOE presente un buen distanciamiento de una clase a otra, este distanciamiento o poder de separación se mide empleando el IV, por lo que aquellas variables con IV alto implican un mayor poder de separación entre clientes buenos y clientes malos.

2.1.3 ANÁLISIS MULTIVARIADO

El último estudio en el análisis característico es el cálculo de la matriz de correlación entre las variables continuas que pasaron previamente los análisis univariado y bivariado. Se realiza con el fin de descartar de acuerdo al nivel de correlación lineal y al grado de multicolinealidad que este posea con el resto de las variables predictivas.

Llamaremos matriz de correlación a la matriz cuadrada y simétrica que tiene unos en la diagonal y fuera de ella los coeficientes de correlación entre las variables. Para este análisis se emplean el coeficiente de correlación de Pearson y el VIF (*Variance Inflation Factor*) o factor de inflación de la varianza:

- **Coeficiente de Correlación de Pearson:** es una medida de la dependencia lineal entre dos variables aleatorias, se calcula como sigue:

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}.$$

  donde $cov(x,y)$ es la covarianza entre x e y y $\sigma_x\sigma_y$ es el producto de las desviaciones estándar de las variables x e y respectivamente.

  El coeficiente de correlación de Pearson es un número real en el rango de $[-1,1]$, una correlación lineal inversa se presenta para valores en $[-1,0)$y una correlación lineal directa se presenta en el intervalo $(0,1]$, por último si $\rho_{xy} = 0$ en general se consideran las variables como independientes.

  Así, para un par de variables con $\rho_{xy} \leq 0.6$ se considerará con una correlación moderada, por lo cual, las variables que no cumplan este parámetro, tendrán una alta correlación y se excluirá aquella que genere menor información para el modelo, es decir, aquella variable que tenga menor IV (criterio definido en el análisis bivariado).

- **VIF *(Variance Inflation Factor)*:** O el factor de inflación de la varianza, se utiliza para medir

el grado de multicolinealidad entre variables:

$$VIF = \frac{1}{1 - R_i^2}.$$

donde $R_i^2 = R^2_{xi|x_1,\ldots,x_{i-1},\ldots,x_{i+1},\ldots,x_k}$, es el coeficiente de correlación múltiple de la variable x_i respecto a las demás variables independientes.

Una regla empírica citada por Kleinbaum [2], consiste en emplear una tolerancia de la siguiente manera:

$$T = \frac{1}{VIF} = 1 - R_i^2.$$

en donde se afirma que existen problemas de colinealidad si el VIF de alguna variable es superior a 10, lo cual corresponde a algún $T < 0.1$ o $R_i^2 > 0.9$.

El proceso de determinación de multicolinealidad es fundamental para eliminar problemas en el modelo general, ya que puede generar una sobreestimación de coeficientes, así como crear valores ficticios sobre los betas de la regresión, llevando a falsas predicciones.

2.2 ANÁLISIS POR ÁRBOLES DE CLASIFICACIÓN

El análisis de cluster es una técnica cuya idea básica es agrupar un conjunto de observaciones en un número dado de *clusters* o grupos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones. La obtención de dichos *clusters* depende del criterio o distancia considerados, es decir, todo depende de lo que consideremos como similar, por lo que es necesario encontrar métodos o algoritmos que infieran el número y componentes de los *clusters* más aceptable, aunque en muchas oportunidades no se obtenga el óptimo absoluto.

Para llevar a cabo la búsqueda de las variables independientes con gran poder predictivo, se uilizará métodos jerárquicos, los cuales actuan en las siguientes dos formas:

- Métodos jerárquicos aglomerativos: se comienza con los objetos o individuos de modo individual; de este modo, se tienen tantos *clusters* iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único *cluster*.

- Métodos jerárquicos divididos: se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén.

Los métodos basados en árboles son métodos jerárquicos bastante populares en *data mining*, pudiéndose usar para clasificación y regresión. Son útiles para la exploración inicial de datos, y apropiados cuando hay un número elevado de datos y existe incertidumbre sobre la manera en que las variables explicativas deberían introducirse en el modelo. Sin embargo, no constituyen una herramienta demasiado precisa de análisis. En conjuntos pequeños de datos es poco probable que revelen la estructura de ellos, de modo que su mejor aplicación se encuentra en grandes masas de datos donde pueden revelar formas complejas en la estructura que no se pueden detectar con los métodos convencionales de regresión.

Los modelos de árbol, donde la variable de destino puede tomar un conjunto finito de valores se denominan árboles de clasificación. En estas estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan las conjunciones de características que conducen a esas etiquetas de clase. Por el contrario, los árboles de decisión, donde la variable de destino puede tomar valores continuos (por lo general números reales) se llaman árboles de regresión. Sin embargo, para ambos árboles el objetivo es crear un modelo que prediga el valor de una variable de destino en función de diversas variables de entrada. Los algoritmos para la construcción de árboles de decisión suelen trabajar de manera *top-down*, escogiendo en cada paso la variable que mejor divide el conjunto de elementos.

Entre otros métodos de minería de datos, los árboles de decisión tienen las siguientes ventajas:

- Facilidad para el entendimiento e interpretación.
- Poca preparación de los datos.
- Capacidad para el manejo tanto de datos numéricos como categorizados.
- Utiliza un modelo de caja blanca.
- Es posible validar su resultado utilizando pruebas estadísticas, haciendo que sea posible tener en cuenta la fiabilidad del modelo.

- Es robusto y es posible realizarse con grandes conjuntos de datos utilizando recursos informáticos estándar en un plazo razonable.

Muchos paquetes de *software* de minería de datos proporcionan implementaciones de uno o varios algoritmos de árboles de decisión. Varios ejemplos incluyen IBM SPSS *Modeler*, *RapidMiner*, SAS *Enterprise Miner*, MATLAB, R, el cual es un entorno de *software* de código abierto para el cálculo estadístico que incluye varias implementaciones tales como los paquetes *rpart*, *party* y *randomForest*; así mismo, se tiene *Weka, Orange*, KNIME, *Microsoft* SQL *Server*, y *scikit-learn* (una biblioteca de aprendizaje automático libre y de código abierto para el lenguaje de programación *Python*).

2.3 REGRESIÓN LOGÍSTICA

Es conocido que la regresión logística es una poderosa herramienta estadística debido a que es un instrumento de análisis bivariado o multivariado de uso tanto explicativo como predictivo que posee la mejor capacidad para analizar datos provenientes de cualquier área en que la variable respuesta sea dicotómica o politómica.

Por sus características, los modelos de regresión logística permiten clasificar individuos dentro de las categorías de la variable dependiente según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables. El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

Para el caso específico de esta investigación se trabajará con regresión logística binaria o binomial. En los modelos de regresión logística binaria la variable dependiente es una variable dicotómica que se codificará como 0 y 1 (cliente bueno y cliente malo respectivamente), es decir, si se considera y_i una variable binaria ($y_i \sim$Bernoulli(p_i)) tal que $y_i = 1$ ocurre con probabilidad p_i definida como $p_i = F(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$ y donde:

- $x_i = (x_1, x_2, ..., x_k)$ es el vector con los valores de k variables explicativas.
- $\beta = (\beta_0, \beta_1, ..., \beta_k)$ es el vector de k coeficientes de regresión.

- F denota una función de distribución acumulada (fda). Para el caso en que F sea la fda de una distribución logística se denotará como $F(t) = \frac{e^t}{1+e^t}$.

La ecuación de partida en los modelos de regresión logística es:

$$P(y=1|x) = p_i = F(\beta_0 + \sum_{i=1}^{k} \beta_i x_i) = \frac{e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}}{1 + e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}}.$$

de donde:

$$p_i + p_i e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)} = e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)},$$

$$p_i = (1 - p_i) e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)},$$

$$\frac{p_i}{(1 - p_i)} = e^{(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}.$$

si ahora se realiza su transformación logaritmo natural, se obtiene una ecuación lineal que es de manejo matemático:

$$\log(\frac{p_i}{1 - p_i}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i.$$

En la expresión anterior se obtiene al lado izquierdo de la igualdad el llamado logit, es decir, el logaritmo natural de la razón de proporciones de éxito y fracaso de la variable dependiente, mientras que el término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal:

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n.$$

Sin embargo la regresión lineal presenta una diferencia fundamental respecto al modelo de regresión logística. En el modelo de regresión lineal se asume que los errores estándar de cada coeficiente siguen una distribución normal de media 0 y varianza constante (homoscedasticidad). En el caso del modelo de regresión logística no pueden realizarse estas asunciones pues la variable dependiente no es continua (sólo puede tomar dos valores, 0 ó 1, pero ningún valor intermedio).

Por lo que si se llama ε al posible error de predicción para cada covariable x_i se tendrá que el error cometido dependerá del valor que llegue a tomar la variable dependiente, es decir:

$$y = P(x) + \varepsilon \qquad \begin{cases} y = 1 & \varepsilon = 1 - P(x). \\ y = 0 & \varepsilon = -P(x). \end{cases}$$

Esto implica que ε sigue una distribución binomial, con media y varianza proporcionales al tamaño muestral y a $P(y = 1|x)$ (la probabilidad de que $y = 1$ dada la presencia de x).

2.3.1 MÉTODO DE SELECCIÓN DE VARIABLES

Para seleccionar las variables se utilizará el método de *stepwise*, en donde se combinan los métodos adelante y atrás, por lo que se puede empezar por el modelo vacío o por el completo, siempre y cuando se mantenga que en cada paso se exploren las variables incluidas, por si deben salir y las no seleccionadas, por si deben entrar.

- **Método hacia adelante:**

  1. Se inicia con un modelo vacío (sólo α).
  2. Se ajusta un modelo y se calcula el p valor de incluir cada variable por separado.
  3. Se selecciona el modelo con la variable más significativa, es decir, aquella que cumpla con los niveles de significancia (α) .
  4. Se ajusta un modelo con la(s) variable(s) seleccionada(s) y se calcula el p valor de añadir cada variable no seleccionada por separado.
  5. Se selecciona el modelo con la variable más significativa, es decir, aquella que cumpla con los niveles de significancia (α).
  6. Se repite 4 y 5 hasta que no queden variables significativas para incluir, es decir, ya no se cuente con variables que cumpla con los niveles de significancia (α).

- **Método hacia atrás:**

  1. Se inicia con un modelo con TODAS las variables candidatas.
  2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar.
  3. Se selecciona para eliminar la variable menos significativa, es decir, aquella variable que no cumpla con los niveles de significancia (α).

4. Se repite 2 y 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

2.4 DESARROLLO DEL MODELO SCORE

Luego de realizar la regresión logística con el conjunto de variables seleccionado al final de cada análisis bajo la metodología de *Stepwise*, se procederá al desarrollo de las *scorecards* de los modelos, es decir, la generación de las fórmulas de puntuación.

El modelo de *score* definitivo cuenta entonces con un conjunto de características, un conjunto de categorías para cada característica y las salidas del modelo de regresión como son el intercepto, los valores beta estimados y estadísticos de rendimiento del modelo (*test* de *Hosmer-Lemeshow*, etc.). A partir de *aquí* se procede a la calibración del modelo. Este proceso relaciona los puntajes obtenidos en el modelo con el logaritmo natural de los cocientes entre Buenos y Malos (odds) y significa relacionar los puntajes obtenidos en el proceso de modelamiento con la probabilidad de incumplimiento que pueden tener los clientes de la institución que es objeto de estudio. Este proceso es importante puesto que ayuda a la entidad bancaria a hacer una comparación de clientes adecuada, tomar decisiones de manera efectiva, entre otros beneficios.

La relación que se obtiene al calibrar los puntajes se puede observar en el siguiente gráfico:

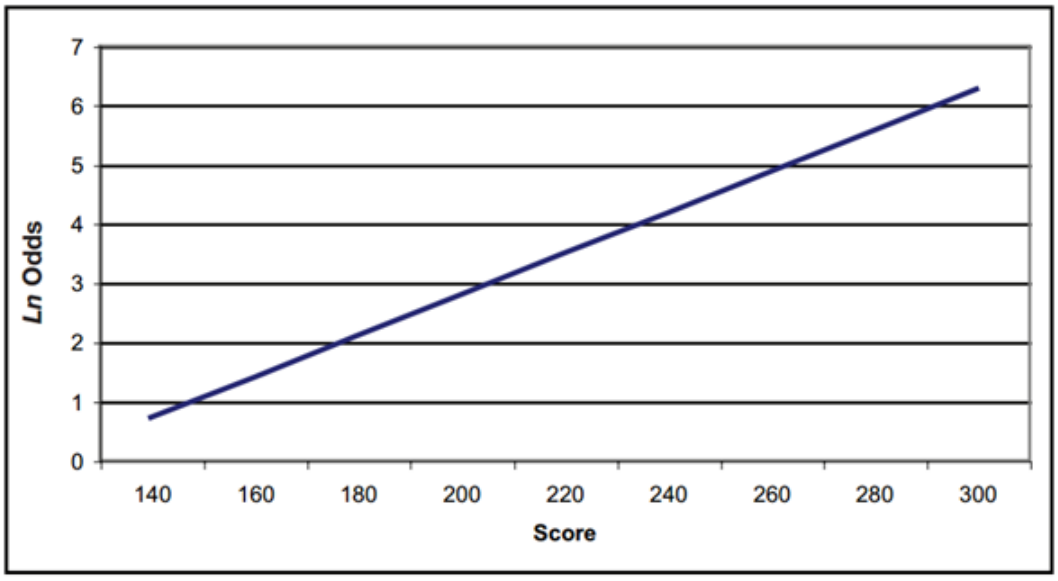

Figura 2.4.1 Relación Ln Odds - *Score*

Se puede observar que los valores de *score* calibrados toman valores discretos. Si bien es cierto que los puntajes se pueden presentar en forma discreta o decimal, se recomienda usar una escala discreta con escala logarítmica. Para realizar el cálculo de los *scores* calibrados, se procede de la siguiente manera:

- Se calcula la Constante de Localización (también conocido como *offset*) mediante la siguiente fórmula:

$$offset = score\ promedio - \left[\frac{PDO}{\ln(2)} * \ln(odds)\right].$$

  donde PDO (*points to double the score*) es un parámetro que mide cuantos puntos cambia el *score* si la relación de *odds* aumenta de 100:1 a 200:1.

- Se calcula el *score* para la variable i en la categoría j:

$$score_{i,j} = \frac{PDO}{\ln(2)} * \beta_{i,j},\ i = 1,\ldots,N;\ j = 1,\ldots,cat(i).$$

- Se procede a calcular el *score* calibrado, de la siguiente manera:

$$score\,calibrado_{i,j} = score_{i,j} + F,\ i = 1,\ldots,N;\ j = 1,\ldots,cat(i).$$

  donde F es $F = \frac{offset}{N+1}$ y N es el número de variables resultantes de la regresión logística. Finalmente, se redondean los *scores* calibrados obtenidos, al entero más próximo.

- Se calculan los pesos por cada categoría para cada variable:

$$w_{ij} = \frac{|\beta_{ij}| * se_{ij}}{\sum_{i=1}^{N-1}\sum_{j=1}^{c(i)} |\beta_{ij}| * se_{ij}},\ i = 1,\ldots,N;\ j = 1,\ldots,c(i).$$

  donde:

  β_{ij}: beta estimado para la variable i en la categoría j.

  $c(i)$: número de categorías de la variable i.

  se_{ij} : error estándar de la estimación para la variable i en la categoría j.

- Por último, se calcula el peso para cada variable categórica:

$$W_i = \sum_{j=1}^{c(i)} w_{ij},\ i = 1,\ldots,N;\ j = 1,\ldots,c(i).$$

2.5 VALIDACIÓN DEL MODELO

La principal herramienta para mirar la predictibilidad de un *score* es realizar una tabla de validación. Una tabla de validación es un cuadro que contiene todos los rangos de *score* resultantes de la etapa de modelación con su distribución de población respectiva manifestando la cantidad de individuos buenos y malos que se presentan por cada rango, así mismo se observa el comportamiento de la tasa de malos a medida que los rangos de *score* disminuyen. Por otra parte las tablas de validación también muestran el valor del estadístico KS y los odds (buenos sobre malos) por cada rango.

Score	Dif	TOTAL			Buenos			Malos				Rel.	Bad/Good	Odds
Range		Dist Int	% Col	%Acum	Dist. Int	% Int.	%Acum	Dist. Int	% Int.	%Acum	KS	% Malos	Odds Acum	Odds
R_8	I_8	n_8	$p_8 = n_8/N$	$pa_8 = pa_7 + p_8$	b_8	$tb_8 = b_8/n_8$	$pb_8 = pb_7 + b_8/B$	m_8	$tm_8 = m_8/n_8$	$pm_8 = pm_7 + m_8/M$	$\lvert pb_8 - pm_8\rvert$	m_8/n_8	b_8/m_8	b_8/m_8
R_7	I_7	n_7	$p_7 = n_7/N$	$pa_7 = pa_6 + p_7$	b_7	$tb_7 = b_7/n_7$	$pb_7 = pb_6 + b_7/B$	m_7	$tm_7 = m_7/n_7$	$pm_7 = pm_6 + m_7/M$	$\lvert pb_7 - pm_7\rvert$	$\Sigma_{i=7}^{8} m_i/n_i$	$\Sigma_{i=7}^{8} b_i/m_i$	b_7/m_7
R_6	I_6	n_6	$p_6 = n_6/N$	$pa_6 = pa_5 + p_6$	b_6	$tb_6 = b_6/n_6$	$pb_6 = pb_5 + b_6/B$	m_6	$tm_6 = m_6/n_6$	$pm_6 = pm_5 + m_6/M$	$\lvert pb_6 - pm_6\rvert$	$\Sigma_{i=6}^{8} m_i/n_i$	$\Sigma_{i=6}^{8} b_i/m_i$	b_6/m_6
R_5	I_5	n_5	$p_5 = n_5/N$	$pa_5 = pa_4 + p_5$	b_5	$tb_5 = b_5/n_5$	$pb_5 = pb_4 + b_5/B$	m_5	$tm_5 = m_5/n_5$	$pm_5 = pm_4 + m_5/M$	$\lvert pb_5 - pm_5\rvert$	$\Sigma_{i=5}^{8} m_i/n_i$	$\Sigma_{i=5}^{8} b_i/m_i$	b_5/m_5
R_4	I_4	n_4	$p_4 = n_4/N$	$pa_4 = pa_3 + p_4$	b_4	$tb_4 = b_4/n_4$	$pb_4 = pb_3 + b_4/B$	m_4	$tm_4 = m_4/n_4$	$pm_4 = pm_3 + m_4/M$	$\lvert pb_4 - pm_4\rvert$	$\Sigma_{i=4}^{8} m_i/n_i$	$\Sigma_{i=4}^{8} b_i/m_i$	b_4/m_4
R_3	I_3	n_3	$p_3 = n_3/N$	$pa_3 = pa_2 + p_3$	b_3	$tb_3 = b_3/n_3$	$pb_3 = pb_2 + b_3/B$	m_3	$tm_3 = m_3/n_3$	$pm_3 = pm_2 + m_3/M$	$\lvert pb_3 - pm_3\rvert$	$\Sigma_{i=3}^{8} m_i/n_i$	$\Sigma_{i=3}^{8} b_i/m_i$	b_3/m_3
R_2	I_2	n_2	$p_2 = n_2/N$	$pa_2 = pa_1 + p_2$	b_2	$tb_2 = b_2/n_2$	$pb_2 = pb_1 + b_2/B$	m_2	$tm_2 = m_2/n_2$	$pm_2 = pm_1 + m_2/M$	$\lvert pb_2 - pm_2\rvert$	$\Sigma_{i=2}^{8} m_i/n_i$	$\Sigma_{i=2}^{8} b_i/m_i$	b_2/m_2
R_1	I_1	n_1	$p_1 = n_1/N$	$pa_1 = p_1$	b_1	$tb_1 = b_1/n_1$	$pb_1 = b_1/B$	m_1	$tm_1 = m_1/n_1$	$pm_1 = m_1/M$	$\lvert pb_1 - pm_1\rvert$	$\Sigma_{i=1}^{8} m_i/n_i$	$\Sigma_{i=1}^{8} b_i/m_i$	b_1/m_1
Total		$N = \Sigma_{i=1}^{8} n_i$	$P = \Sigma_{i=1}^{8} p_i$		$B = \Sigma_{i=1}^{8} b_i$	TTB = B/N		$M = \Sigma_{i=1}^{8} m_i$	TTM = M/N		$\max\lvert pb_i - pm_i\rvert$			B/M

Figura 2.5.1 Ejemplo Tabla de Validación

En dichas tablas se encontrarán:

- **Rango del *score*:** es la cantidad de intervalos en las cuáles se dividirá los puntajes de *score*; para la construcción de los intervalos, se utilizan cuantiles que dividan de manera adecuada al conjunto de *scores* obtenidos del modelo de regresión, según el número de intervalos que sean requeridos. Es decir, si se quieren construir 10 rangos de *score*, se deben utilizar los deciles del 1 al 10 para dividir al conjunto de *scores*, con el fin de hallar granularidad para mejorar la toma de decisiones.

- **Amplitud de intervalo (Dif):** definido como la distancia que existe entre el *score* máximo y el *score* mínimo dentro de cada rango.

- **Total de la población:** indica cual es el total de población en el intervalo R_i (Dist Int), así como el porcentaje que representa con respecto a toda la población (%Col) y el porcentaje acumulado de población en el intervalo R_i.(%Acum).

- **Total de buenos:** indica cual es el total de buenos en el intervalo R_i (Dist Int), así como el porcentaje que representa con respecto a toda la población (%Int) y el porcentaje acumulado de buenos en el intervalo R_i (%Acum). Es de esperar que la tasa de buenos aumente a mayor *score* obtenido, es decir, la tasa de buenos es creciente.

- **Total de malos:** indica cual es el total de malos en el intervalo R_i (Dist Int), así como el porcentaje que representa con respecto a toda la población (%Int) y el porcentaje acumulado de malos en el intervalo R_i (%Acum). Es de esperar que la tasa de malos disminuya a mayor *score* obtenido, es decir, la tasa de malos es decreciente.

- **Odd:** es el cociente de buenos entre malos en el intervalo R_i, es de esperar que a mayor *score* obtenido el cociente sea mayor, es decir, el Odds es creciente. Su fórmula viene dada por:

$$Odds_i = {}^{\#Buenos_i}/_{\#Malos_i}.$$

  Si se quiere conocer de manera relativa, la fórmula es la siguiente:

$$Odds_i = {}^{\%Buenos_i}/_{\%Malos_i}.$$

- **KS:** se define como la mayor diferencia entre las distribuciones acumuladas de buenos y malos $max|pb_i - pm_i|$ se calcula en base a las probabilidades de *default* y no *default*, el KS varía entre el rango 0 – 100 y es deseable que su valor se encuentre entre 20 – 70. Este *test* se utiliza para probar las siguientes hipótesis:

  – H0: la distribución del *score* de las cuentas buenas es la misma que la distribución del *score* de las cuentas malas.

  – H1: la distribución del *score* de las cuentas buenas no es la misma que la distribución del *score* de las cuentas malas.

  Se rechaza H0 si $Dmax > Dcrítico$ donde $Dcrítico = 100 * K\sqrt{\frac{1}{B}+\frac{1}{M}}$.

  – B = Total de cuentas buenas.

  – M = Total de cuentas malas.

  – K = Valor obtenido de la Tabla de *Kolmogorov-Smirnov* asociado a un nivel de significación en particular.

En general, para validar el modelo a través de la tabla de validación, es necesario verificar lo siguiente:

- Predicción: para establecer el nivel de predicción la principal medida es el KS. La estadística de *Kolmogorov – Smirnov* (KS) es usada para determinar el nivel de separación del *score*; esta es una prueba que busca analizar las semejanzas entre dos distribuciones, sin embargo, es adaptada para el caso del *score* en donde no se busca semejanza, sino diferencia entre las distribuciones de los clientes que son considerados buenos y aquellos considerados como malos. Los valores de KS deben estar dentro de los valores señalados y entre mayor sea el KS mayor es el nivel de predicción del *score*.

- Ordenamiento: es importante prestar especial atención a dos columnas, la primera es el porcentaje de los clientes denominados malos, con la finalidad de observar que a mayor puntaje del *score*, menor porcentaje de clientes malos se encuentren en el rango. La segunda columna que debe observarse es la de odds la cual debe estar ordenada de mayor a menor en el mismo sentido que el *score*, es decir, entre mayor sea el *score* mayor proporción de odds debe tener.

  En el ordenamiento debe observarse que no existan saltos, es decir, que se mantenga un comportamiento siempre monótono creciente o decreciente. Para el caso en los que se tienen menos de 6 rangos en el *score* se deben incluir más variables en el modelo ya que esto indicaría que las variables incluidas no son suficientes para discriminar entre buenos y malos.

- Granularidad en el porcentaje total: el porcentaje total por rangos debe estar distribuido uniformemente, para garantizar granularidad en la población. El porcentaje mínimo permitido por rango será de 5%, si esto no ocurre se debe agrupar el *score* en menos rangos.

2.6 INFERENCIA DE RECHAZADOS

Para que un modelo de *score* sea representativo en toda la población a la que se va a evaluar, debe poder identificarse solicitudes que fueron previamente aprobadas pero que deberían haber sido rechazadas y, además, debe poder identificar solicitudes que fueron previamente rechazadas

que deberían haber sido aprobadas. Dado que es imposible conocer el comportamiento de los rechazados con la entidad, se realiza inferencia sobre estos.

La inferencia de rechazados es el proceso de estimación del riesgo de incumplimiento de los nuevos solicitantes que resultan denegados bajo las políticas de evaluación actuales de una entidad. La importancia de este proceso radica en el sesgo que puede producirse si se implementara un modelo de *scoring* ignorando los denegados, dado que su comportamiento podría diferir en gran medida del de la población de aprobados, con lo cual el *score* no sería aplicable a toda la población de nuevos solicitantes.

2.6.1 INFERENCIA DE RECHAZO POR EL MÉTODO DE *PARCELING*

Existen diversas metodologías para aplicar inferencia de rechazados, siendo una de las más precisas la metodología *parceling*. Esta metodología consiste en construir un modelo de Bueno/Malo a partir de la población conocida (población aprobada). En función de este modelo, se asignan las probabilidades de ser malo en la población de rechazados.

La metodología *Parceling*, se compone de las siguientes etapas:

1. **Modelo de inferencia**: cálculo de la probabilidad de *default* (PD) y desarrollo del modelo de inferencia con población aceptada (*Known Good/Bad Model*).
2. **Calificación con el modelo de inferencia:** luego del desarrollo del modelo de la población aprobada, se deberá inferir el comportamiento de los denegados. Para ello se implementa el modelo de inferencia en la población de denegados.
3. **Inferencia del comportamiento de los rechazados:** se define el comportamiento de los clientes Bueno/Malo en la población de rechazados, bajo números aleatorios y la probabilidad de ser malo.

Por ejemplo, se tiene la siguiente PD (tasa de malos) y *score* obtenidos del modelo con la población aprobada (paso 1):

Score	**Malos Aprobados**	**Buenos Aprobados**	**Tasa de Malos (TM)**
+250	139	6811	2%
210 - 249	213	4047	5%
⋮	⋮	⋮	⋮
170 - 189	530	2414	18%
130 - 169	290	971	23%

Tabla 2.6.1.1 Ejemplo para realizar la Inferencia de Rechazo con el Método de *Parceling*

Luego, se implementará dicho modelo en la población rechazada (paso 2) para inferir si su comportamiento es bueno o malo (paso 3); así cada cliente rechazado se le será asignado un *score* y un número aleatorio $k \sim \mathscr{U}(0,1)$, donde se definirá la variable BGI de los rechazados de la siguiente manera:

$$BGI_{Reject} = \begin{cases} 1 & si\, k \leq 2 * TM_{rango}. \\ 0 & si\, k > 2 * TM_{rango}. \end{cases}$$

Es decir, si un cliente rechazado obtuvo luego de la implementación un *score* de 220 puntos, y por otra parte se le fue asignado un valor aleatorio k = 0,5, el cliente se le considerará un cliente bueno (BGI = 0) ya que $k > 2 * TM_{rango} \Rightarrow 0.5 > 2 * 0,05 = 0.1$.

CAPÍTULO 3

METODOLOGÍA

Este trabajo consiste en el desarrollo de un modelo de *scoring* para los clientes nuevos de una entidad argentina. En términos generales, las etapas consideradas en dicha investigación se resumen en:

- Tratamiento de la base de datos.
- Definición de la población objetivo y variable respuesta.
- Desarrollo y evaluación univariada, bivariada y multivariada de las variables predictivas.
- Desarrollo de los árboles de clasificación de las variables predictivas.
- Modelado: desarrollo del *score*, análisis de su estabilidad y capacidad de discriminación para la población aprobada.
- Inferencia de rechazados.
- Modelado: desarrollo del *score* final para la población aprobada y rechazada.

Mediante el desarrollo de las etapas descritas se obtiene un modelo de *score* que permite optimizar el proceso de admisión de clientes y gestionar la cartera de forma más adecuada. Estos modelos, a su vez, consideraran durante el proceso de desarrollo, información del Buró de Crédito, información demográficas e información interna de la entidad. Toda la información se encuentra estructurada lo que garantiza integridad y consistencia en la información y todos los análisis se llevará a cabo con el *software* estadístico R.

3.1 TRATAMIENTO DE LA BASE DE DATOS

Luego de la recepción de los archivos de datos requeridos para el desarrollo de este trabajo de investigación por parte de la entidad bancaria, se realizó un análisis cuantitativo y cualitativo de los mismos con el fin de garantizar la validez de la información proporcionada. Se verificó que el 100% de las variables no presentaran ningún inconveniente ni complicación al momento de la lectura y procesamiento. Se evidenciaron tablas de frecuencia para variables categóricas, revisión

particular de las variables en formato fecha y las métricas estándar para las variables continuas. Los periodos de observación enviados corresponden a los meses comprendidos entre julio 2012 y septiembre 2013. Por lo que se concluye que la misma se encuentra apta para ser utilizada en el desarrollo del modelo.

Para la elaboración de los modelos de *scoring*, fue necesario efectuar exclusiones de registros y transformaciones a las variables que componen la base de datos suministrada por el cliente. Las exclusiones de esta población se estableció por temas de negocio, para conservar la homogeneidad de la información y evitar los sesgos de población atípica.

- Clientes con 30 o menos días de Mora y con 90 o más días de Mora: el modelo de *score* desarrollado en este trabajo sólo puede ser aplicado para aquellas cuentas con estado de mora menor a 30 y mayor o igual a 90 días, procurando realizar una gestión únicamente con clientes buenos y malos, definido más adelante. No se recomienda su uso para cuentas con estados de moras entre 30 y 90 días, dado que no se realizan validación alguna sobre los clientes que presentan esta morosidad.

- Edad: los clientes con edad superior o igual a 99 años deben ser excluidos del cálculo del *score*. En particular al estudiar la base enviada por la entidad bancaria se encontró un caso de un cliente con esta edad, el cual fue excluido del análisis.

- Datos Atípicos: en la variable CLTE_Actividad_Empresa se encontraron valores iguales a "?", específicamente se encontraron 77.982 casos, por otro lado, la variable ALTATC_Canal-_Origen presentó 2 casos que señalaban "SIN INFORMAR"; los casos de ambas variables fueron imputados a *missing* evitando así los sesgos por población atípica.

- Clientes con valores *missing*: se evidenció en la variable CLTE_Segmento la presencia de clientes con información faltante, en particular se encontraron 36 casos. Esta casuística coincidió a su vez con las variables: CLTE_Actividad_persona, CLTE_Estado_civil, CLTE_Actividad_Empresa, CLTE_Sucursal, CLTE_provincia, CLTE_Convenio y CLTE-_Edad, que presentaron información carente en los mismos 36 casos anteriores; dichos casos fueron excluídos de la base.

- Imputación de *missing*: las variables numéricas que sean predictivas para el modelo, se imputarán los *missing* (NA) a el valor -99.

Por último, las variables de saldo, movimientos en cuentas, y demás variables que estaban expresadas en miles de pesos argentinos, se deflactaron para llevarlas a cantidad de sueldos en base al sueldo mínimo vigente, es decir, si un cliente tiene como saldo 6.600 ARS en agosto del 2013, su saldo deflactado corresponderá a 2 sueldos mínimos, como se muestra en la siguiente tabla:

Desde	Valor bruto en moneda local	Variación nominal	Variación interanual	Valor en USD[1]	Valor en USD constantes (agosto de 1993)[2]	Normativa
2011-09	2.300 ARS	25%	32,18%	547,61	349,07	Res. 2/11 CNEPySMVyM
2012-09	2.670 ARS	16,09%	16,09%	569,29	355,70	Res. 2/12 CNEPySMVyM
2013-02	2.850 ARS	6,74%	16,67%	570,00	356,34	Res. 2/12 CNEPySMVyM
2013-08	3.300 ARS	15,79%	23,91%	583,03	360,10	Res. 4/13 CNEPySMVyM
2014-01	3.600 ARS	9,09%	34,83%	450,00	340,59	Res. 4/13 CNEPySMVyM

Figura 3.1.1 Tabla de Sueldos Mínimos en Argentina en el período septiembre 2011 - enero 2014

3.2 DEFINICIÓN DE LA POBLACIÓN OBJETIVO Y VARIABLE RESPUESTA

La variable respuesta BGI de sus siglas *Bad Good and Indeterminate*; permite estudiar el evento de estudio y su determinación de acuerdo a la altura de morosidad (90 o más días de mora) y tiempo de maduración (12 meses). La altura de morosidad corresponde a los días vencidos que pueden ser manejados mediante los vectores de mora de cada cliente. Para los modelos que se realizarán se contará con una ventana de períodos desde septiembre 2012 hasta septiembre 2013.

Para este trabajo se creó la variable respuesta llamada BGI (*Bad Good and Indeterminate*) de la siguiente manera:

- -4: sin comportamiento (no tiene comportamiento en ninguno de los 12 meses posteriores al punto de observación).
- -3: comportamiento insuficiente (tiene comportamiento en menos de 6 meses posteriores al punto de observación).
- -1: indeterminado (la máxima mora en los 12 meses posteriores al punto de observación es >= 30 y < 90).
- 0: bueno (la máxima mora en los 12 meses posteriores al punto de observación es < 30).
- 1: Malo (la máxima mora en los 12 meses posteriores al punto de observación es >= 90).

Sin embargo, para realizar cualquiera de los dos modelos se tomaron en cuenta únicamente aquellos registros que presenten BGI 0 y 1, es decir, los registros definidos como buenos y malos, tal como se especificó en la sección de exclusión anterior.

3.3 ANÁLISIS CARACTERÍSTICO

El análisis de cada característica, implica entender su comportamiento de forma continua o categórica y su relación con respecto al evento a modelar. El aporte al modelo de cada característica se evalúa de manera individual y colectiva. Para este proceso se realizan tres etapas: análisis univariado, bivariado y multivariado.

3.3.1 ANÁLISIS UNIVARIADO

El análisis univariado evalúa las variables desde el punto de vista de su distribución, variabilidad y completitud para realizar un primer filtro y encontrar las mejores variables explicativas. Los parámetros en esta fase hacen referencia a la completitud y variabilidad.

- Completitud: cada una de las variables a ser consideradas como explicativas deben ser completas en por lo menos un 5%, es decir, el total de elementos faltantes (*missing*) no supera o iguala al 95%, si es así se consideraran incompletas y serán excluidas del análisis.
- Variabilidad: cada una de las variables explicativas disponibles para desarrollar el modelo deben ser lo suficientemente variables para que sean potencialmente útiles para discriminar

las categorías de la variable respuesta. Cada una de las variables debe tener una variabilidad mínima del 2% para poder ser considerada en el conjunto de variables explicativas, es decir, si el P_1 y P_{99} son diferentes entre sí.

- Cota Superior: cada una de las variables explicativas se les calcula el P_{95} y todos los valores que presente la variable mayores a él, se acotaron a este valor para evitar así valores atípicos.
- Cota Inferior: cada una de las variables explicativas se les calcula el P_5 y todos los valores que presente la variable menores a él, se acotaron a este valor para evitar así valores atípicos.

En resumen, se tendrá:

Parámetros	**Valor**
Completitud	0.05
Variabilidad	P2
Cota Superior	P95
Cota Inferior	P5

Tabla 3.3.1.1 Parámetros fijados por el cliente para el Análisis Univariado

3.3.2 ANÁLISIS BIVARIADO

Este análisis consiste en evaluar los tramos óptimos de las variables categóricas y continuas, para asegurar que los mismos representen un porcentaje importante de la población, en donde se verifica lo siguiente: el *InformationValue* (IV) obtenido sea significativo (mayor a 0.02) y la distribución del *badrate* sea monótona (ordenamiento lógico), es decir, evaluar que el resultado esperado de tasa de malos concuerde con el sentido lógico de la variable. Esta comprobación se realiza ya que se puede tener variables contra intuitivas que a pesar de tener buen IV no van acorde con el evento de estudio o viceversa, variables con buen sentido lógico pero con IV muy pequeño.

3.3.3 ANÁLISIS MULTIVARIADO

En el análisis multivariado se calculan las correlaciones entre todas las variables continuas que cumplieron los criterios de los análisis univariado y bivariado. Esto se hace con el fin de prever futuros problemas de colinealidad. El coeficiente de correlación de Pearson debe ser menor a 0.6, para concluir baja correlación. Otra forma de observar correlación dentro de un modelo es a través del Índice de Inflación de la varianza (VIF). El estándar es que este valor sea inferior a 7 y solo aplica para variables continuas.

3.4 ÁRBOLES DE CLASIFICACIÓN

La clasificación mediante el ánalisis de árboles se implementó con ayuda de la librería *party* la cual contiene un conjunto de funciones para realizar particionamiento recursivo, en la que se encuentra *ctree*. *CTree* es una clase no paramétrica de árboles de regresión dentro de una teoría bien definida de procedimientos de inferencia condicional. Es aplicable a todo tipo de problemas de regresión, incluidos los nominales, ordinales, variables de respuesta numéricas y multivariadas, así mismo es útil por la flexibilidad y herramientas computacionales para adaptar y visualizar este tipo de árboles de inferencia condicional.

Al momento de clasificar se buscó que en cada nodo se ubicara mínimo 5% de la población total para lograr así, una distribución homogénea de los registros de la base de datos. Por otra parte, aquellas variables que resultaron significativas para la clasificación, se tomaron como las variables con mayor poder predictivo para explicar el modelo de regresión logística.

3.5 MODELADO DE LA POBLACIÓN APROBADA

Luego de superados los pasos anteriores, se procedió al desarrollo de las *scorecards* de los modelos, es decir, la generación de las fórmulas de puntuación. La metodología considerada para el desarrollo es mediante análisis de regresión logística ya que ofrece beneficios en interpretación, seguimiento e implementación.

Se realizaron dos regresiones logísticas, una con el conjunto de variables seleccionado al final del análisis univariado, bivariado y multivariado, y otra con el conjunto de variables obtenidas por la clasificación de los árboles, ambas bajo la metodología de *Stepwise*, donde se verificó:

- Significancia: las variables deben ser estadísticamente significativas a un nivel de 5%.
- Ordenamiento tasa de malos: las variables deben presentar orden en la tasa de malos a través de los rangos del *score*.
- Interpretación coherente al negocio: el orden de la tasa de malos, y su correspondiente beta a través de las categorías de las variables categóricas y continuas, no deben tener una interpretación contra intuitiva al contexto de negocio.

- Signo esperado: el signo de los betas estimados debe ser igual al signo de betas intuitivo en el análisis bivariado.
- Número de parámetros/variables: luego de realizar los procedimientos necesarios para obtener las mejores variables explicativas, es necesario garantizar que se tiene un mínimo de 15 parámetros y/o 6 variables en el modelo final.
- Número de categorías: se consideraron variables óptimas aquellas que tuvieron por los menos 2 categorías más el *missing*.
- Índice de condición (colinealidad): el índice de condición de las variables del modelo no deben superar el valor de 7 según sea la tipología del proyecto. Si se presenta un índice de condición mayor es necesario quitar la variable del modelo e incluir una que pueda explicar de forma similar el fenómeno representado.

Para el desarrollo del *score* se realiza la alineación, o calibración en donde se define según el modelo, el puntaje donde se encuentra un malo por cada bueno, y a partir de ahí cuando se aumente el puntaje en 80 puntos se doble la cantidad de odds. Generalmente este puntaje es de 600 y la cantidad que aumenta es 80.

Por último se construye la tabla de validación que contiene rangos del *score* contra los posibles valores de la variable respuesta, es decir buenos y malos y en donde verificamos:

- Ordenamiento: verificar que el porcentaje de malos ordene a través de los rangos del *score*. Si esto no ocurre se debe agrupar el *score* en menos rangos.
- Número de rangos: si se tienen menos de 6 rangos en el *score* se deben incluir más variables en el modelo ya que esto indicaría que las variables incluidas no son suficientes para discriminar entre buenos y malos.

3.6 INFERENCIA DE RECHAZADOS

Luego del desarrollo del modelo sobre la población aprobada, se realizó la inferencia del comportamiento de la población rechazada mediante la metodología *Parceling*. Para ello se realizan los siguientes pasos:

- Estimar la probabilidad de ser malo en la población de rechazados implementando el modelo obtenido con la población aprobada.

- Asignar a cada registro rechazado un número aleatorio con el comando *runif* en R.
- Definir un cliente bueno/malo en la población de rechazados basándose en el parámetro que la tasa de malos de la población rechazada es 2 veces peor que la de la población aceptada, es decir, si un cliente presenta un número aleatorio menor o igual a 2 veces la tasa de malos con respecto a la población aceptada, este cliente se considerará malo, de lo contrario, será un cliente bueno.

3.7 MODELO DE *SCORE* FINAL

El modelo final se desarrolló con la base de datos de clientes aprobados y rechazados con la nueva variable respuesta (BGI), que correspondía al comportamiento de bueno/malo inferido. Dicha base se sometió nuevamente al análisis característico y árboles de clasificación (según fue el caso) para la obtención de las nuevas variables predictivas que se utilizaron en la regresión logística. Asimismo, se llevaron a cabo pruebas para determinar la estabilidad y calibración de los modelos generados y se obtuvieron los indicadores de desempeño del *score*, que aportaron de manera significativa en la discriminación de los clientes considerados como buenos y malos.

CAPÍTULO 4

ANÁLISIS DE RESULTADOS

Con los datos suministrados por la entidad bancaria argentina, se planteó hacer una comparación entre la escogencia de variables por parte del análisis característico y los árboles de clasificación para conocer qué conjunto de variables predicen mejor un modelo de regresión logística. Dicha comparación se realizó tanto para el modelo con la población aprobada como para el modelo final con toda la población (aprobada + rechazada). Estos análisis fueron implementados usando la plataforma R *Studio*.

Para mostrar los resultados de los modelos tanto con la selección de variables obtenidas mediante análisis característico y árboles de clasificación, se comparará el valor del AIC, KS y la significancia entre los dos grupos de variables que resultaron predictivas dentro del modelo.

4.1 RESULTADOS OBTENIDOS SOBRE LA POBLACIÓN APROBADA

A continuación se mostrarán los resultados tanto del análisis característico como el de los árboles de clasificación que se llevaron a cabo para seleccionar las variables predictivas, así mismo los modelos y sus respectivas validaciones que se realizaron con la población de aprobados.

4.1.1 RESULTADOS OBTENIDOS DEL ANÁLISIS CARACTERÍSTICO

Para realizar el análisis característico se cuenta con una base de datos de 124.277 registros correspondiente a los clientes aprobados por la entidad bancaria y 150 variables en las que se encuentra información demográfica, financiera y de buró de crédito de los clientes.

4.1.1.1 RESULTADOS DEL ANÁLISIS UNIVARIADO

Para el análisis univariado, primero se acotaron las 150 variables al P_5 y el P_{95}, esto quiere decir que todos los valores que estaban por debajo del P_5 pasaron a tomar este valor y todos los valores que eran mayores al P_{95}, tendrán el valor correspondiente a su P_{95}. Una vez acotada las variables,

se estudió la completitud al 5%, que no es más que el porcentaje mínimo requerido de información (valores distintos a NA) que debe presentar cada variable. Y finalmente se calculó la variabilidad del 2%, en general se obtuvo que:

Variables	Frecuencia	Frecuencia Relativa
Excluidas	36	24%
Aceptadas	114	76%

Tabla 4.1.1.1.1 Resumen de los resultados obtenidos en el Análisis Univariado

En relación al desgloce por variables se obtuvo los siguientes resultados:

Variables	% Información	Completitud	P1	P99	Variabilidad	Uso
VERAZ_VERSION	53,28%	Aceptar	1	1	Excluir	Excluir
HP_Sexo	14,43%	Aceptar	1	2	Aceptar	Aceptar
VERAZ_PRE_PLAZO_NOACTIVAS	0%	Excluir	0	0	Excluir	Excluir
CLTE_provincia	99,97%	Aceptar	1	25	Aceptar	Aceptar
ALTAPP_Monto_deflac	2,71%	Excluir	2,46	22,11	Aceptar	Excluir

Tabla 4.1.1.1.2 Ejemplo de los resultados obtenidos en el Análisis Univariado

4.1.1.2 RESULTADOS DEL ANÁLISIS BIVARIADO

Una vez obtenido los resultados del análisis univariado, las 114 variables que cumplieron con los parámetros de completitud y variabilidad, fueron sometidas al análisis bivariado. Se procedió a categorizar cada variable para calcular posteriormente el IV, el cual debe ser mayor a 0.02 para considerarse predictiva. Obteniendo los siguientes resultados.

Variables	IV	Uso
VERAZ_AUTOMATICO	0,030	Aceptar
CLTE_Actividad_Empresa	0,255	Aceptar
ALTATC_Forma_Pago	0,015	Excluir
VERAZ_TAR_LIM_MAX	0,117	Aceptar
ALTATC_Sucursal	0,014	Excluir

Tabla 4.1.1.2.1 Ejemplo de los resultados obtenidos en el Análisis Bivariado

En general se obtuvo:

Variables	Frecuencia	Frecuencia Relativa
Predictivas	60	52,63%
No Predictivas	54	47,37%

Tabla 4.1.1.2.2 Resumen de los resultados obtenidos en el Análisis Bivariado

4.1.1.3 RESULTADOS DEL ANÁLISIS MULTIVARIADO

Una vez obtenida las 60 variables que resultaron predictivas del análisis bivariado, se sigue a ubicar cuáles son variables numéricas y cuáles son variables categóricas.

Variables Predictivas	Frecuencia	Frecuencia Relativa
Numéricas	42	70%
Categóricas	18	30%

Tabla 4.1.1.3.1 Proporción de Variables Numéricas y Categóricas

A las 42 variables numéricas se le procedió a calcular la matriz de correlación y el cálculo del VIF. Para el análisis de correlación se tomó un parámetro del 60% mientras que para el VIF se aceptaron todas aquellas variables cuyo valor resultó ser menor a 7. Así, se obtuvo que 34 variables se extrayeron por correlación, resultando 8 variables numéricas aptas para el modelado. Al cacular el VIF de estas 8 variables, todos resultaron menores a 7.

Las variables que superaron el análisis multivariado son:

Variables	VIF
ALTATC_Limite_Compra	2,2088
CLTE_Edad	1,4951
VERAZ_TAR_MESES_ANTIGUEDAD	2,0222
VERAZ_TAR_PAGOS_MIN_deflac	1,6447
VERAZ_CUOTA	1,5575
VERAZ_CTA_ACDO_ACTIVAS	1,2841
HP_FLT_Ing_Inferido	1,2952
Calificación_Concepto_Modif	1,1276

Tabla 4.1.1.3.2 Variables que resultaron no Correlacionadas

Así, por la metodología del análisis característico, de 150 variables con las que se inició el análisis, 26 variables resultaron aptas o predictivas (8 variables numéricas resultantes del análisis multivariado y 18 variables categóricas resultantes del análisis bivariado) para ser utilizadas en la regresión logística.

4.1.2 RESULTADOS OBTENIDOS DE LOS ÁRBOLES DE CLASIFICACIÓN

Para realizar el análisis con árboles de clasificación se contó con la misma base de datos de 124.277 registros correspondiente a los clientes aprobados y 150 variables en las que se encuentra información demográfica, financiera y de buró de crédito de los clientes. Para realizar el árbol se utilizó el comando *ctree,* y se buscó que en cada nodo se ubicara mínimo 5% de la población total para lograr así, una distribución homogénea de los registros de la base de datos. Obteniéndose el siguiente árbol:

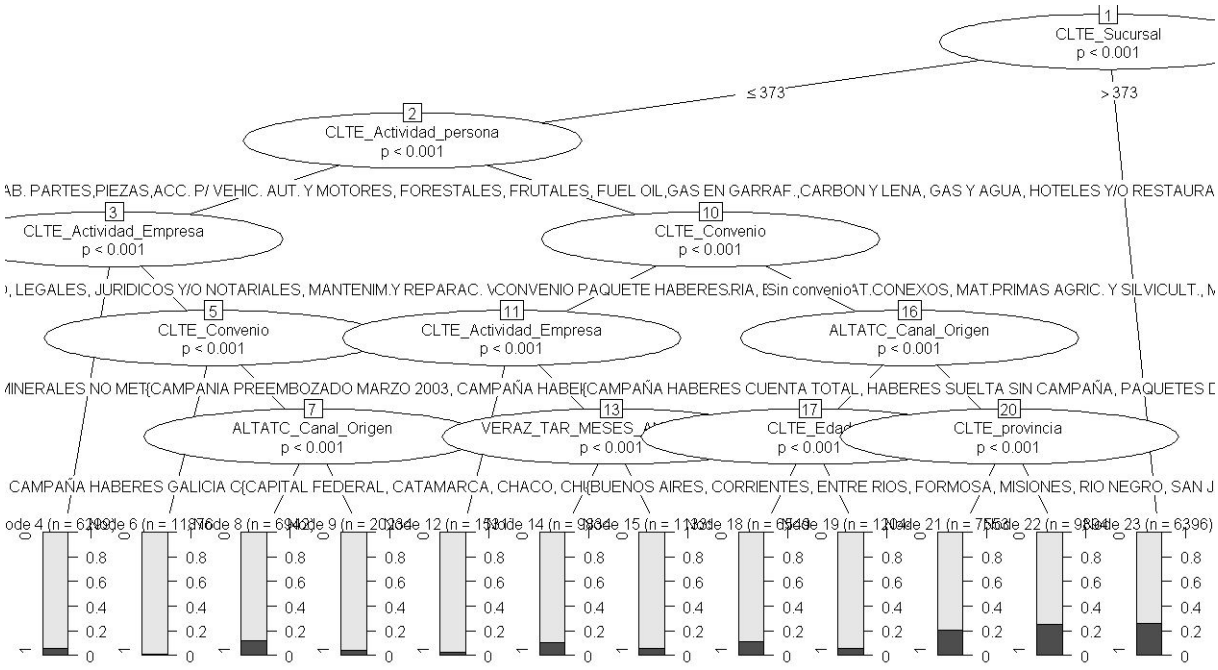

Figura 4.1.2.1 Árbol de Clasificación Obtenido

En resumen, por la metodología de árboles se obtuvo que 8 variables resultaron aptas o predictivas para ser utilizadas en la regresión logística. Las cuales se muestran con detalle a continuación:

Variables	Nodos	Naturaleza
CLTE_Sucursal	1	Categórica
CLTE_Actividad_persona	2	Categórica
CLTE_Actividad_Empresa	3-11	Categórica
CLTE_Convenio	5-10	Categórica
ALTATC_Canal_Origen	7-16	Categórica
VERAZ_TAR_MESES_ANTIGUEDAD	13	Numérica
CLTE_Edad	17	Numérica
CLTE_provincia	20	Categórica

Tabla 4.1.2.1 Variables que resultaron Aptas por el árbol de Clasificación

4.1.3 RESULTADOS OBTENIDOS EN LA REGRESIÓN LOGÍSTICA

Para realizar la regresión logística se utilizaron únicamente las variables que resultaron aptas de los análisis explicados anteriormente y el 70% de los registros totales de la base de datos, es decir, 86.994 registros, los cuales fueron tomados aleatoriamente.

4.1.3.1 MODELO - ANÁLISIS CARACTERÍSTICO

En esta sección se mostrará el modelo producto de las 26 variables que resultaron aptas luego que cumplieran los parámetros mínimos requerido en los análisis univariado, bivariado y multivariado.

Se realizaron 6 versiones de modelo en donde se fueron depurando las variables que no eran significativas ($p - valor < 0.05$) resultando el siguiente:

Variable	Categoría	Coef	Std. Error	Pr(>|z|)	Peso
Intercepto		-0,532	0,062	< 2e-16	
ALTATC_Alcance	Platinum, Corporate, Signature, Oro	0,411	0,061	1.93e-11	3,12%
CALIF_Estado_Calificacion_PP	Missing, Calificado, No disponible no malo	0,197	0,048	3.35e-05	1.17%
CALIF_Producto_Generador	Missing	0,154	0,029	1.82e-07	0,56%
CLTE_Actividad_Empresa	Joyeria y/o fantasia, Trasp. automotor,...	0,976	0,034	< 2e-16	31,8%
	Servicio Agricola, Servicio consultoria,...	1,147	0,046	< 2e-16	
	Servicios Técnicos, Servicios Forestales,...	1,726	0,098	< 2e-16	

CLTE_Actividad_persona	Fab. de Calzado y/o sus partes,...	0,280	0,043	8.88e-11	3,58%
	Energia electrica, Elab Cacao y Chocolate,...	0,432	0,039	< 2e-16	
CLTE_Edad	24 < CLTE_Edad ≤38	0,148	0,037	5.93e-05	10,4%
	38 < CLTE_Edad ≤43	0,286	0,054	9.39e-08	
	43 < CLTE_Edad ≤52	0,397	0,053	8.92e-14	
	52 < CLTE_Edad ≤Inf	0,728	0,057	< 2e-16	
CLTE_provincia	Tucuman, Mendoza, Cordoba, Buenos Aires	0,112	0,030	0.00023	3,97%
	No Informada, La Pampa, Capital Federal	0,644	0,044	< 2e-16	
CLTE_Estado_civil	Missing, Viudo, Separado, Divorciado,...	0,492	0,051	< 2e-16	3,10%
CLTE_Segmento	2110050	0,561	0,035	< 2e-16	17.3%
	1230020, 2110010, 2110040	0,683	0,033	< 2e-16	
	Missing, 2120010, 2110080, 2110030,...	1,468	0,066	< 2e-16	
CLTE_Sucursal	Missing, 976, 909, 782, 484, 433, 377,...	0,771	0,029	< 2e-16	13.2%
	997, 927, 827, 667, 439, 379, 371, 363,...	0,870	0,041	< 2e-16	
	990, 911, 825, 500, 437, 378, 370, 360,...	1,098	0,045	< 2e-16	
VERAZ_TAR_MESES _ANTIGUEDAD	19 < VERAZ_TAR_MESES_ANTIGUEDAD ≤36	0,463	0,066	1.72e-12	11,8%
	36 < VERAZ_TAR_MESES_ANTIGUEDAD ≤80	0,524	0,052	< 2e-16	
	80 < VERAZ_TAR_MESES_ANTIGUEDAD ≤Inf	0,603	0,062	< 2e-16	

Tabla 4.1.3.1.1 Modelo de Regresión Logística - Análisis Característico

En general, se tiene un modelo con 11 variables de las cuales 2 son numéricas y 9 son variables categóricas. Todas son significativas ya que su p-valor es < 0.05 y sus coeficientes presentan monotonía con respecto a la definición de las variables. Los estadísticos que sustenta el modelo anterior son:

Estadístico	**Valor**
AIC	46163
Residual Deviance	46113
Grados de Libertad (R.D)	86969
Null Deviance	53716
Grados de Libertad (N.D)	86993

Tabla 4.1.3.1.2 Estadísticos del Modelo - Análisis Característico

4.1.3.2 MODELO - ÁRBOLES DE CLASIFICACIÓN

En esta sección se mostrará el modelo producto de las 8 variables que resultaron aptas del árbol de clasificación. Se realizaron 4 versiones de modelo en donde se fueron depurando las variables que no eran significativas ($p - valor < 0.05$) resultando el siguiente:

Variable	Categoría	Coef	Std. Error	Pr(>|z|)	Peso
Intercepto		0,348	0,048	4.23e-13	
CLTE_Actividad_persona	Fab. de Calzado y/o sus partes,...	0,335	0,042	2.21e-15	5,49%
	Energia electrica, Elab Cacao y Chocolate,...	0,593	0,038	< 2e-16	
CLTE_Actividad_Empresa	Joyeria y/o fantasia, Trasp. automotor,...	1,097	0,033	< 2e-16	43,37%
	Servicio Agricola, Servicio consultoria,...	1,340	0,045	< 2e-16	
	Servicios Técnicos, Servicios Forestales,...	1,954	0,097	< 2e-16	
CLTE_Edad	24 < CLTE_Edad ≤38	0,460	0,033	< 2e-16	26,02%
	38 < CLTE_Edad ≤43	0,799	0,049	< 2e-16	
	43 < CLTE_Edad ≤52	0,977	0,048	< 2e-16	
	52 < CLTE_Edad ≤Inf	1,386	0,052	< 2e-16	
VERAZ_TAR_MESES _ANTIGUEDAD	19 < VERAZ_TAR_MESES_ANTIGUEDAD ≤36	0,508	0,064	3.14e-15	18,2%
	36 < VERAZ_TAR_MESES_ANTIGUEDAD ≤80	0,663	0,051	< 2e-16	
	80 < VERAZ_TAR_MESES_ANTIGUEDAD ≤Inf	0,893	0,060	< 2e-16	
CLTE_provincia	Tucuman, Mendoza, Cordoba, Buenos Aires	0,232	0,029	8.28e-16	6,97%
	No Informada, La Pampa, Capital Federal	0,936	0,042	< 2e-16	

Tabla 4.1.3.2.1 Modelo de Regresión Logística - Árboles de Clasificación

En general, se tiene un modelo con 5 variables de las cuales 2 son numéricas y 3 son variables categóricas. Todas son significativas ya que su p-valor es < 0.05 y sus coeficientes presentan monotonía con respecto a la definición de las variables. Los estadísticos que sustenta el modelo anterior son:

Estadístico	Valor
AIC	48889
Residual Deviance	48859
Grados de Libertad (R.D)	86979
Null Deviance	53716
Grados de Libertad (N.D)	86993

Tabla 4.1.3.2.2 Estadísticos del Modelo - Árboles de Clasificación

4.1.4 RESULTADOS OBTENIDOS DE LA VALIDACIÓN DE LOS MODELOS

Para verificar los modelos de regresión logística se calculó inicialmente el *score* para cada registro, luego se generaron tablas de validación para las poblaciones de entrenamiento (70% utilizado para realizar la regresión), de validación (30% restante) y total (100% de la población) y en donde se calculó entre otros valores, el KS, Odds y tasa de malos para cada rango de *score*.

4.1.4.1 TABLAS DE VALIDACIÓN - MODELO ANÁLISIS CARACTERÍSTICO

A continuación se muestran las tablas obtenidas del modelo con el análisis característico:

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
732	992	8520	9,79	8433	98,98	87	1,02	0	96,93
680	731	8615	9,9	8442	97,99	173	2,01	9,6	48,8
643	679	8796	10,11	8503	96,67	293	3,33	18,16	29,02
612	642	8647	9,94	8290	95,87	357	4,13	25,29	23,22
589	611	8520	9,79	8031	94,26	489	5,74	31,38	16,42
562	588	8582	9,87	8014	93,38	568	6,62	35,48	14,11
531	561	9063	10,42	8343	92,06	720	7,94	38,59	11,59
502	530	8468	9,73	7524	88,85	944	11,15	40,23	7,97
458	501	9023	10,37	7473	82,82	1550	17,18	38,06	4,82
302	457	8760	10,07	5877	67,09	2883	32,91	28,3	2,04
Total		**86994**	**100**	**78930**	**90,73**	**8064**	**9,27**	**40,23**	**9,79**

Figura 4.1.4.1.1 Tabla de Entrenamiento para el Modelo Aprobado - Análisis Característico

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
731	979	3664	9,83	3620	98,8	44	1,2	0	82,27
679	730	3673	9,85	3598	97,96	75	2,04	9,44	47,97
643	678	3665	9,83	3556	97,03	109	2,97	17,92	32,62
610	642	3778	10,13	3621	95,84	157	4,16	25,31	23,06
587	609	3725	9,99	3497	93,88	228	6,12	31,49	15,34
560	586	3716	9,97	3471	93,41	245	6,59	35,27	14,17
530	559	3670	9,84	3382	92,15	288	7,85	38,49	11,74
500	529	3793	10,17	3334	87,9	459	12,1	40,19	7,26
457	499	3743	10,04	3118	83,3	625	16,7	36,84	4,99
302	456	3856	10,34	2613	67,76	1243	32,24	28,06	2,1
Total		**37283**	**100**	**33810**	**90,68**	**3473**	**9,32**	**40,19**	**9,74**

Figura 4.1.4.1.2 Tabla de Validación para el Modelo Aprobado - Análisis Característico

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
732	992	12116	9,75	11987	98,94	129	1,06	0	92,92
680	731	12280	9,88	12031	97,97	249	2,03	9,51	48,32
643	679	12537	10,09	12134	96,79	403	3,21	18,02	30,11
611	642	12442	10,01	11930	95,88	512	4,12	25,3	23,3
588	610	12342	9,93	11613	94,09	729	5,91	31,44	15,93
561	587	12435	10,01	11623	93,47	812	6,53	35,42	14,31
531	560	12433	10	11439	92,01	994	7,99	38,7	11,51
501	530	12441	10,01	11012	88,51	1429	11,49	40,23	7,71
458	500	12629	10,16	10477	82,96	2152	17,04	37,6	4,87
302	457	12622	10,16	8494	67,3	4128	32,7	28,25	2,06
Total		**124277**	**100**	**112740**	**90,72**	**11537**	**9,28**	**40,23**	**9,77**

Figura 4.1.4.1.3 Tabla Total para el Modelo Aprobado - Análisis Característico

Las tablas anteriores presentaron una apertura de 10 rangos en el *score* lo cual garantiza granularidad en el estudio, así mismo, presentaron un comportamiento monótono creciente en la tasa de malos y un comportamiento monótono decreciente en los odds lo que coincide con la coherencia de la clasificación de los rangos (a mayor rango de *score*, menor tasa de malos y mayor odds; y a menor rango de *score*, mayor tasa de malos y menor odds). En relación al KS, se obtuvo un valor del 40,23% con lo cual se considera una buena discriminación de la variable respuesta.

4.1.4.2 TABLAS DE VALIDACIÓN - MODELO ÁRBOLES DE CLASIFICACIÓN

En esta sección se mostrará las tablas de entrenamiento, validación y total obtenidas del modelo por la metodología de los árboles de clasificación:

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
668	879	9451	10,86	9280	98,19	171	1,81	0	54,27
632	667	9738	11,19	9445	96,99	293	3,01	9,64	32,24
607	631	8335	9,58	7946	95,33	389	4,67	17,97	20,43
586	606	7533	8,66	7122	94,54	411	5,46	23,21	17,33
563	585	12305	14,14	11472	93,23	833	6,77	27,14	13,77
536	562	10247	11,78	9422	91,95	825	8,05	31,35	11,42
501	535	9928	11,41	8863	89,27	1065	10,73	33,05	8,32
471	500	8822	10,14	7398	83,86	1424	16,14	31,07	5,2
378	470	10635	12,22	7982	75,05	2653	24,95	22,79	3,01
Total		**86994**	**100**	**78930**	**90,73**	**8064**	**9,27**	**33,05**	**9,79**

Figura 4.1.4.2.1 Tabla de Entrenamiento para el Modelo Aprobado - Árboles de Clasificación

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
668	879	4006	10,74	3921	97,88	85	2,12	0	46,13
632	667	4173	11,19	4058	97,24	115	2,76	9,15	35,29
607	631	3612	9,69	3440	95,24	172	4,76	17,84	20
584	606	4664	12,51	4364	93,57	300	6,43	23,06	14,55
562	583	4250	11,4	3971	93,44	279	6,56	27,33	14,23
536	561	3769	10,11	3467	91,99	302	8,01	31,05	11,48
501	535	4376	11,74	3904	89,21	472	10,79	32,6	8,27
471	500	3787	10,16	3176	83,87	611	16,13	30,56	5,2
378	470	4646	12,46	3509	75,53	1137	24,47	22,36	3,09
Total		**37283**	**100**	**33810**	**90,68**	**3473**	**9,32**	**32,6**	**9,74**

Figura 4.1.4.2.2 Tabla de Validación para el Modelo Aprobado - Árboles de Clasificación

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
668	879	13457	10,83	13201	98,1	256	1,9	0	51,57
632	667	13911	11,19	13503	97,07	408	2,93	9,49	33,1
607	631	11947	9,61	11386	95,3	561	4,7	17,93	20,3
584	606	15909	12,8	14920	93,78	989	6,22	23,17	15,09
563	583	12211	9,83	11424	93,55	787	6,45	27,83	14,52
536	562	14648	11,79	13474	91,99	1174	8,01	31,14	11,48
501	535	14304	11,51	12767	89,25	1537	10,75	32,91	8,31
471	500	12609	10,15	10574	83,86	2035	16,14	30,92	5,2
378	470	15281	12,3	11491	75,2	3790	24,8	22,66	3,03
Total		**124277**	**100**	**112740**	**90,72**	**11537**	**9,28**	**32,91**	**9,77**

Figura 4.1.4.2.3 Tabla Total para el Modelo Aprobado - Árboles de Clasificación

Las tablas anteriores presentaron una apertura de 9 rangos en el *score* lo cual garantiza granularidad en el estudio, así mismo, presentaron un comportamiento monótono creciente en la tasa de malos y un comportamiento monótono decreciente en los odds lo que coincide con la coherencia de la clasificación de los rangos (a mayor rango de *score*, menor tasa de malos y mayor odds; y a menor rango de *score*, mayor tasa de malos y menor odds). En relación al KS, se obtuvo un valor del 32,91% con lo cual se considera una buena discriminación de la variable respuesta.

4.2 INFERENCIA DE RECHAZO

Para completar el modelo de *score* se identificaron las solicitudes que fueron previamente rechazadas por el cliente y se le realizó inferencia para conocer su comportamiento.

4.2.1 INFERENCIA DE RECHAZO - MODELO ANÁLISIS CARACTERÍSTICO

Luego de realizar el modelo sobre la población de aprobados con la metodología del análisis característico, se prosiguió a inferir sus rechazados por el método de *Parceling*, obteniéndose lo siguiente:

Población	Registros	Buenos	Malos	Tasa de Malos
Aprobada	124.277	112.740	11.537	9,28%
Rechazada	4.537	4.346	191	4,21%
Total	128.814	117.086	11.728	9,1%

Tabla 4.2.1.1 Distribución Total de Buenos y Malos - Análisis Característico

De lo anterior se puede comentar que se presenta discrepancia a nivel negocio ya que se hubiese esperado que la tasa de malos de los rechazados superara a la de los clientes aprobados, sin embargo, basándonos en los objetivos planteados en este trabajo, lo consideraremos como resultado favorable para continuar con la comparación entre las metodologías de análisis característico y árboles de clasificación.

4.2.2 INFERENCIA DE RECHAZO - MODELO ÁRBOLES DE CLASIFICACIÓN

Manteniendo la idea del ítem anterior, sobre la población de aprobados, se realizó un segundo modelo con la metodología de árboles de clasificación, una vez obtenido, se prosiguió a inferir sus rechazados por el método de *Parceling*, obteniéndose lo siguiente:

Población	Registros	Buenos	Malos	Tasa de Malos
Aprobada	124.277	112.740	11.537	9,28%
Rechazada	4.537	4.366	171	3,77%
Total	128.814	117.106	11.708	9,09%

Tabla 4.2.2.1 Distribución Total de Buenos y Malos - Árboles de Clasificación

De lo anterior se puede comentar que se presenta discrepancia a nivel negocio ya que se hubiese esperado que la tasa de malos de los rechazados superara a la de los clientes aprobados, sin embargo, basándonos en los objetivos planteados en este trabajo, lo consideraremos como resultado favorable para continuar con la comparación entre las metodologías de análisis característico y árboles de clasificación.

4.3 RESULTADOS OBTENIDOS SOBRE LA POBLACIÓN TOTAL

Una vez obtenido el comportamiendo de todos los clientes, se continúa con el modelado final manteniendo la metodología utilizada desde el inicio, es decir, se tendrá un modelo final para el modelo que se construyó desde el análisis característico y otro modelo raiz de lo hallado en los árboles de clasificación.

4.3.1 RESULTADOS OBTENIDOS DEL ANÁLISIS CARACTERÍSTICO

Para realizar el análisis característico se cuenta con una base de datos de 128.814 registros correspondiente a todos los clientes (aprobados más rechazados) de la entidad bancaria y 150 variables en las que se encuentra información demográfica, financiera y de buró de crédito de los clientes.

4.3.1.1 RESULTADOS DEL ANÁLISIS UNIVARIADO

Para el análisis univariado, como en el caso de la población aprobada, se inició acotando las 150 variables al P_5 y el P_{95}, esto quiere decir que todos los valores que estaban por debajo del P_5 pasaron a tomar este valor y todos los valores que eran mayores al P_{95}, tendrán el valor correspondiente a su P_{95}. Una vez acotada las variables, se estudió la completitud al 5%, que no es más que el porcentaje mínimo requerido de información (valores distintos a NA) que debe presentar cada variable. Y finalmente se calculó la variabilidad del 2%, en general se obtuvo que:

Variables	Frecuencia	Frecuencia Relativa
Excluidas	33	22%
Aceptadas	117	78%

Tabla 4.3.1.1.1 Resumen de los resultados obtenidos en el Análisis Univariado

En relación al desgloce por variables se obtuvo los siguientes resultados:

Variables	% Información	Completitud	P1	P99	Variabilidad	Uso
VERAZ_VERSION	52,97%	Aceptar	1	1	Excluir	Excluir
HP_Sexo	14,94%	Aceptar	1	2	Aceptar	Aceptar
VERAZ_PRE_PLAZO_NOACTIVAS	0%	Excluir	0	0	Excluir	Excluir
CLTE_provincia	99,97%	Aceptar	1	25	Aceptar	Aceptar
ALTAPP_Monto_deflac	3,17%	Excluir	2,46	22,48	Aceptar	Excluir

Tabla 4.3.1.1.2 Ejemplo de los resultados obtenidos en el Análisis Univariado

4.3.1.2 RESULTADOS DEL ANÁLISIS BIVARIADO

Una vez obtenido los resultados del análisis univariado, las 117 variables que cumplieron con los parámetros de completitud y variabilidad, fueron sometidas al análisis bivariado. Se procedió a categorizar cada variable para calcular posteriormente el IV, el cual debe ser mayor a 0,02 para considerarse predictiva. Obteniendo los siguientes resultados.

Variables	IV	Uso
VERAZ_AUTOMATICO	0,022	Aceptar
CLTE_Actividad_Empresa	0,17	Aceptar
ALTATC_Forma_Pago	0,013	Excluir
VERAZ_TAR_LIM_MAX	0,118	Aceptar
ALTATC_Sucursal	0,023	Aceptar

Tabla 4.3.1.2.1 Ejemplo de los resultados obtenidos en el Análisis Bivariado

En general se obtuvo:

Variables	Frecuencia	Frecuencia Relativa
Predictivas	57	48,72%
No Predictivas	60	51,28%

Tabla 4.3.1.2.2 Resumen de los resultados obtenidos en el Análisis Bivariado

4.3.1.3 RESULTADOS DEL ANÁLISIS MULTIVARIADO

Una vez obtenida las 57 variables que resultaron predictivas del análisis bivariado, se sigue a ubicar cuáles son variables numéricas y cuáles son variables categóricas.

Variables Predictivas	Frecuencia	Frecuencia Relativa
Numéricas	42	73,68%
Categóricas	15	26,32%

Tabla 4.3.1.3.1 Proporción de Variables Numéricas y Categóricas

A las 42 variables numéricas se le procedió a calcular la matriz de correlación y el cálculo del VIF. Para el análisis de correlación se tomó un parámetro del 60% mientras que para el VIF se aceptaron todas aquellas variables cuyo valor resultó ser menor a 7. Así, se obtuvo que 34 variables se extrajeron por correlación, resultando 8 variables numéricas aptas para el modelado. Al cacular el VIF de estas 8 variables, todos resultaron menores a 7.

Las variables que superaron el análisis multivariado son:

Variables	VIF
ALTATC_Limite_Compra	2,1343
CLTE_Edad	1,4855
VERAZ_TAR_MESES_ANTIGUEDAD	2,0124
VERAZ_TAR_PAGOS_MIN_deflac	1,6527
VERAZ_CUOTA_deflac	1,5929
VERAZ_CTA_ACDO_ACTIVAS	1,3007
HP_FLT_Ing_Inferido	1,2727
Calificación_Concepto_Modif	1,1091

Tabla 4.3.1.3.2 Variables que resultaron no Correlacionadas

Así, por la metodología del análisis característico, de 150 variables con las que se inició el análisis, 23 variables resultaron aptas o predictivas (8 variables numéricas resultantes del análisis multivariado y 15 variables categóricas resultantes del análisis bivariado) para ser utilizadas en la regresión logística.

4.3.2 RESULTADOS OBTENIDOS DE LOS ÁRBOLES DE CLASIFICACIÓN

Para realizar el análisis con árboles de clasificación se contó con la misma base de datos de 128.814 registros y 150 variables correspondiente a la información de todos los clientes (aprobados

y rechazados). Para realizar el árbol se utilizó nuevamente el comando *ctree,* y se buscó que en cada nodo se ubicara mínimo 5% de la población total para lograr así, una distribución homogénea de los datos. Obteniéndose:

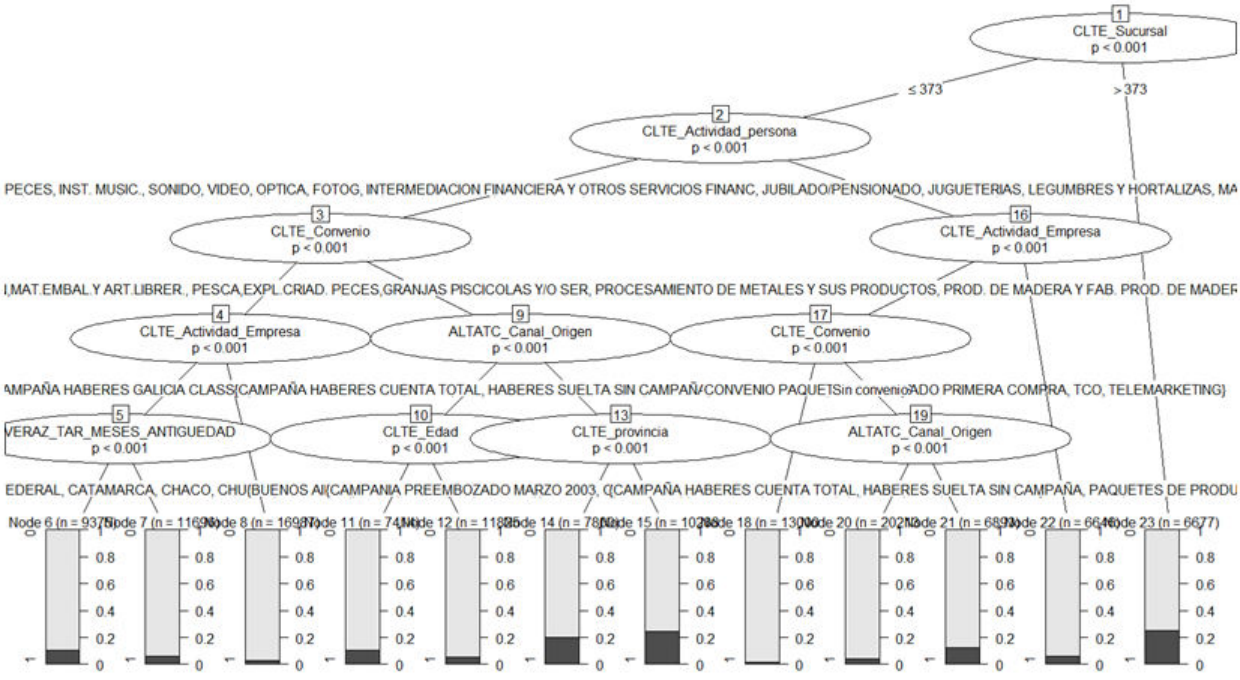

Figura 4.3.2.1 Árbol de Clasificación Obtenido

En resumen, por la metodología de árboles se obtiene que 8 variables resultaron predictivas para ser utilizadas en la regresión logística. Las cuales se muestran a continuación:

Variables	Nodos	Naturaleza
CLTE_Sucursal	1	Categórica
CLTE_Actividad_persona	2	Categórica
CLTE_Convenio	3-17	Categórica
CLTE_Actividad_Empresa	4-16	Categórica
VERAZ_TAR_MESES_ANTIGUEDAD	5	Numérica
ALTATC_Canal_Origen	9-19	Categórica
CLTE_Edad	10	Numérica
CLTE_provincia	13	Categórica

Tabla 4.3.2.1 Variables que resultaron Aptas por el árbol de Clasificación

4.3.3 RESULTADOS OBTENIDOS EN LA REGRESIÓN LOGÍSTICA

Para realizar la regresión logística se utilizó únicamente las variables que resultaron aptas de los análisis explicados anteriormente y el 70% de los registros totales de la base de datos, es decir, 90.170 registros, los cuales fueron tomados aleatoriamente.

4.3.3.1 MODELO - ANÁLISIS CARACTERÍSTICO

En esta sección se mostrará el modelo producto de las 23 variables que resultaron aptas luego que cumplieran los parámetros mínimos requerido en los análisis univariado, bivariado y multivariado.

Se realizaron 9 versiones de modelo en donde se fueron depurando las variables que no eran significativas ($p - valor < 0.05$) resultando el siguiente:

Variable	Categoría	Coef	Std. Error	Pr(>|z|)	Peso
Intercepto		-0,141	0,060	0,018	
CLTE_Segmento	2110100, 1220020, 2110050	0,608	0,034	< 2e-16	25,6%
	1230020, 2110010, 2110040	0,724	0,032	< 2e-16	
	2120010, 2110080, 1220040, 1230030 ...	1,439	0,064	< 2e-16	
ALTATC_Alcance	Platinum, Corporate, Signature, Oro	0,393	0,059	3,51e-11	4,37%
CLTE_Estado_civil	Viudo, Separado, Divorciado, Concubino...	0.482	0,049	< 2e-16	4,45%
CLTE_provincia	Tucuman, Cordoba, Buenos Aires	0,136	0,029	0.,000002	5,61%
	No Informada, La Pampa, Capital Federal	0,614	0,042	< 2e-16	
ALTATC_Sucursal	380, 369, 351, 313, 250, 199, 165, 100,...	0,308	0,038	1,13e-15	9,23%
	375, 366, 338, 300, 233, 197, 157, 83,...	0,431	0,039	< 2e-16	
	373, 363, 332, 274, 225, 181, 149, 82,...	0,677	0,031	< 2e-16	
VERAZ_AUTOMATICO	Si	0,133	0,028	0.,000001	0,69%
CLTE_Actividad_Empresa	Hoteles y/o Restaurantes, Jugueterias,...	0,974	0,041	< 2e-16	14,8%
	Joyeria y/o Fantasia, Enseñanza,...	1,180	0,033	< 2e-16	
CLTE_Actividad_persona	Transporte Aereo, Ama de Casa...	0,133	0,045	0.,0032	6,05%
	Florerias, Muebles y Colchones,...	0,438	0,060	2,84e-13	
CALIF_Producto_Generador	Missing	0,236	0,029	2,5e-16	5,11%
	SC sin antiguedad	0,387	0,053	2,68e-13	

CLTE_Edad	25 < CLTE_Edad ≤38	0,233	0,031	1,49e-13	14,5%
	38 < CLTE_Edad ≤43	0,300	0,048	6,33e-10	
	43 < CLTE_Edad ≤52	0,382	0,048	248e-15	
	52 < CLTE_Edad ≤Inf	0,704	0,053	< 2e-16	
VERAZ_TAR_MESES_ANTIGUEDAD	19 < VERAZ_TAR_MESES_ANTIGUEDAD ≤60	0,418	0,051	2,74e-16	9,54%
	60 < VERAZ_TAR_MESES_ANTIGUEDAD ≤Inf	0,569	0,052	< 2e-16	

Tabla 4.3.3.1.1 Modelo de Regresión Logística - Análisis Característico

En general, se tiene un modelo con 11 variables de las cuales 2 son numéricas y 9 son variables categóricas. Todas son significativas ya que su p-valor es < 0.05 y sus coeficientes presentan monotonía con respecto a la definición de las variables. Los estadísticos que sustenta el modelo anterior y servirán para compararlo con el siguiente modelo son:

Estadístico	Valor
AIC	48198
Residual Deviance	48150
Grados de Libertad (R.D)	90146
Null Deviance	55190
Grados de Libertad (N.D)	90169

Tabla 4.3.3.1.2 Estadísticos del Modelo - Análisis Característico

4.3.3.2 MODELO - ÁRBOLES DE CLASIFICACIÓN

En esta sección se mostrará el modelo producto de las 8 variables que resultaron aptas del árbol de clasificación. Se realizaron 2 versiones de modelo en donde se fueron depurando las variables que no eran significativas ($p - valor < 0.05$) resultando el siguiente:

Variable	Categoría	Coef	Std. Error	Pr(>|z|)	Peso
Intercepto		0,386	0,054	6,05e-13	
CLTE_Sucursal	976, 927, 856, 782, 667, 460, 433...	0,572	0,032	< 2e-16	15,83%
	1125, 960, 911, 852, 767, 500...	0,695	0,035	< 2e-16	
	997, 945, 909, 827, 706, 463, 437...	0,846	0,035	< 2e-16	
CLTE_Actividad_persona	Transporte Aereo, Ama de Casa...	0,106	0,045	0,018619	9,00%
	Florerias, Muebles y Colchones,...	0,606	0,059	< 2e-16	
CLTE_Actividad_Empresa	Hoteles y/o Restaurantes, Jugueterias,...	1,077	0,041	< 2e-16	19,03%
	Joyeria y/o Fantasia, Enseñanza,...	1,312	0,032	< 2e-16	

VERAZ_TAR_MESES	19 < VERAZ_TAR_MESES_ANTIGUEDAD ≤60	0,574	0,048	< 2e-16	14,88%
_ANTIGUEDAD	60 < VERAZ_TAR_MESES_ANTIGUEDAD ≤Inf	0,829	0,048	< 2e-16	
CLTE_Edad	25 < CLTE_Edad ≤38	0,486	0,030	< 2e-16	34,75%
	38 < CLTE_Edad ≤43	0,760	0,046	< 2e-16	
	43 < CLTE_Edad ≤52	0,929	0,045	< 2e-16	
	52 < CLTE_Edad ≤Inf	1,326	0,049	< 2e-16	
CLTE_provincia	Tucuman, Cordoba, Buenos Aires	0,099	0,028	0,000449	6,51%
	No Informada, La Pampa, Capital Federal	0,647	0,041	< 2e-16	

Tabla 4.3.3.2.1 Modelo de Regresión Logística - Árboles de Clasificación

En general, se tiene un modelo con 6 variables de las cuales 2 son numéricas y 4 son variables categóricas. Todas son significativas ya que su p-valor es < 0.05 y sus coeficientes presentan monotonía con respecto a la definición de las variables. Los estadísticos que sustenta el modelo anterior son:

Estadístico	**Valor**
AIC	49611
Residual Deviance	49579
Grados de Libertad (R.D)	90154
Null Deviance	55190
Grados de Libertad (N.D)	90169

Tabla 4.3.3.2.2 Estadísticos del Modelo - Árboles de Clasificación

4.3.4 RESULTADOS OBTENIDOS DE LA VALIDACIÓN DE LOS MODELOS

Para verificar los modelos de regresión logística se calculó inicialmente el *score* para cada registro, luego se generaron tablas de validación para las poblaciones de entrenamiento (70% utilizado para realizar la regresión), de validación (30% restante) y total (100% de la población) y en donde se calculó entre otros valores, el KS, Odds y tasa de malos para cada rango de *score*.

4.3.4.1 TABLAS DE VALIDACIÓN - MODELO ANÁLISIS CARACTERÍSTICO

A continuación se muestran las tablas obtenidas del modelo con el análisis característico:

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
720	932	8878	9,85	8786	98,96	92	1,04	0	95,5
670	719	8913	9,88	8732	97,97	181	2,03	9,6	48,24
637	669	8950	9,93	8670	96,87	280	3,13	18,04	30,96
608	636	8946	9,92	8554	95,62	392	4,38	25,21	21,82
585	607	9078	10,07	8566	94,36	512	5,64	30,87	16,73
561	584	8913	9,88	8293	93,04	620	6,96	35,08	13,38
534	560	9040	10,03	8232	91,06	808	8,94	37,64	10,19
503	533	8857	9,82	7866	88,81	991	11,19	37,84	7,94
461	502	9420	10,45	7921	84,09	1499	15,91	35,36	5,28
336	460	9175	10,18	6345	69,16	2830	30,84	26,75	2,24
Total		**90170**	**100**	**81965**	**90,9**	**8205**	**9,1**	**37,84**	**9,99**

Figura 4.3.4.1.1 Tabla de Entrenamiento para el Modelo Final - Análisis Característico

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
720	932	3906	10,11	3865	98,95	41	1,05	0	94,27
670	719	3868	10,01	3774	97,57	94	2,43	9,84	40,15
637	669	3927	10,16	3805	96,89	122	3,11	17,92	31,19
608	636	3841	9,94	3684	95,91	157	4,09	25,29	23,46
585	607	3909	10,12	3671	93,91	238	6,09	31,32	15,42
561	584	3725	9,64	3473	93,23	252	6,77	35,02	13,78
534	560	3900	10,09	3530	90,51	370	9,49	37,76	9,54
503	533	3784	9,79	3343	88,35	441	11,65	37,31	7,58
461	502	3927	10,16	3275	83,4	652	16,6	34,3	5,02
336	460	3857	9,98	2701	70,03	1156	29,97	25,12	2,34
Total		**38644**	**100**	**35121**	**90,88**	**3523**	**9,12**	**37,76**	**9,97**

Figura 4.3.4.1.2 Tabla de Validación para el Modelo Final - Análisis Característico

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
720	932	12784	9,92	12651	98,96	133	1,04	0	95,12
670	719	12781	9,92	12506	97,85	275	2,15	9,67	45,48
637	669	12877	10	12475	96,88	402	3,12	18,01	31,03
608	636	12787	9,93	12238	95,71	549	4,29	25,23	22,29
585	607	12987	10,08	12237	94,22	750	5,78	31	16,32
561	584	12638	9,81	11766	93,1	872	6,9	35,06	13,49
534	560	12940	10,05	11762	90,9	1178	9,1	37,67	9,98
503	533	12641	9,81	11209	88,67	1432	11,33	37,68	7,83
461	502	13347	10,36	11196	83,88	2151	16,12	35,04	5,21
336	460	13032	10,12	9046	69,41	3986	30,59	26,26	2,27
Total		**128814**	**100**	**117086**	**90,9**	**11728**	**9,1**	**37,68**	**9,98**

Figura 4.3.4.1.3 Tabla Total para el Modelo Final - Análisis Característico

Las tablas anteriores presentaron una apertura de 10 rangos en el *score* lo cual garantiza granularidad en el estudio, así mismo, presentaron un comportamiento monótono creciente en la tasa de malos y un comportamiento monótono decreciente en los odds lo que coincide con la coherencia de la clasificación de los rangos (a mayor rango de *score*, menor tasa de malos y mayor odds; y a menor rango de *score*, mayor tasa de malos y menor odds). En relación al KS, se obtuvo un valor del 37,68% con lo cual se considera una buena discriminación de la variable respuesta.

4.3.4.2 TABLAS DE VALIDACIÓN - MODELO ÁRBOLES DE CLASIFICACIÓN

En esta sección se mostrará las tablas de entrenamiento, validación y total obtenidas del modelo por la metodología de los árboles de clasificación:

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
681	865	8941	9,92	8768	98,07	173	1,93	0	50,68
648	680	8819	9,78	8572	97,2	247	2,8	8,59	34,7
622	647	8888	9,86	8522	95,88	366	4,12	16,02	23,28
601	621	9083	10,07	8620	94,9	463	5,1	21,95	18,62
579	600	9218	10,22	8706	94,45	512	5,55	26,81	17
558	578	7855	8,71	7356	93,65	499	6,35	31,17	14,74
534	557	10205	11,32	9359	91,71	846	8,29	34,05	11,06
507	533	8714	9,66	7733	88,74	981	11,26	35,14	7,88
468	506	9311	10,33	7822	84,01	1489	15,99	32,59	5,25
383	467	9136	10,13	6524	71,41	2612	28,59	23,94	2,5
Total		**90170**	**100**	**81982**	**90,92**	**8188**	**9,08**	**35,14**	**10,01**

Figura 4.3.4.2.1 Tabla de Entrenamiento para el Modelo Final - Árboles de Clasificación

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
681	865	3906	10,11	3821	97,82	85	2,18	0	44,95
648	680	3870	10,01	3750	96,9	120	3,1	8,47	31,25
622	647	3742	9,68	3601	96,23	141	3,77	15,74	25,54
601	621	3905	10,11	3718	95,21	187	4,79	21,98	19,88
579	600	3889	10,06	3659	94,09	230	5,91	27,25	15,91
558	578	3455	8,94	3222	93,26	233	6,74	31,13	13,83
534	557	4356	11,27	4002	91,87	354	8,13	33,68	11,31
507	533	3744	9,69	3341	89,24	403	10,76	35,03	8,29
468	506	3945	10,21	3285	83,27	660	16,73	33,09	4,98
383	467	3832	9,92	2725	71,11	1107	28,89	23,69	2,46
Total		**38644**	**100**	**35124**	**90,89**	**3520**	**9,11**	**35,03**	**9,98**

Figura 4.3.4.2.2 Tabla de Validación para el Modelo Final - Árboles de Clasificación

Min Score	Max Score	Total	% Total	Buenos	Tasa Buenos	Malos	Tasa Malos	KS	Odds
681	865	12847	9,97	12589	97,99	258	2,01	0	48,79
648	680	12689	9,85	12322	97,11	367	2,89	8,55	33,57
622	647	12630	9,8	12123	95,99	507	4,01	15,93	23,91
601	621	12988	10,08	12338	95	650	5	21,95	18,98
579	600	13107	10,18	12365	94,34	742	5,66	26,94	16,66
558	578	11310	8,78	10578	93,53	732	6,47	31,16	14,45
534	557	14561	11,3	13361	91,76	1200	8,24	33,94	11,13
507	533	12458	9,67	11074	88,89	1384	11,11	35,1	8
468	506	13256	10,29	11107	83,79	2149	16,21	32,74	5,17
383	467	12968	10,07	9249	71,32	3719	28,68	23,86	2,49
Total		**128814**	**100**	**117106**	**90,91**	**11708**	**9,09**	**35,1**	**10**

Figura 4.3.4.2.3 Tabla Total para el Modelo Final - Árboles de Clasificación

Las tablas anteriores presentaron una apertura de 10 rangos en el *score* lo cual garantiza granularidad en el estudio, así mismo, presentaron un comportamiento monótono creciente en la tasa de malos y un comportamiento monótono decreciente en los odds lo que coincide con la coherencia de la clasificación de los rangos (a mayor rango de *score*, menor tasa de malos y mayor odds; y a menor rango de *score*, mayor tasa de malos y menor odds). En relación al KS, se obtuvo un valor del 35,1% con lo cual se considera una buena discriminación de la variable respuesta.

CONCLUSIONES Y RECOMENDACIONES

En el presente trabajo de grado se compararon dos metodologías para seleccionar variables predictivas que se utilizarían en una posterior regresión logística, se implementó con el *software* R y se tomó como base los datos suministrados por una entidad bancaria argentina. En relación a las diferentes etapas en las que se realizó este proyecto, se pudo concluir lo siguiente:

ANÁLISIS SOBRE LA POBLACIÓN APROBADA

- **Selección de Variables:** para la selección de variables se utilizó el análisis característico y los árboles de clasificación, en donde se obtuvo que por el primer método señalado resultaron 26 variables aptas mientras que por el análisis de árboles se obtuvieron 8 variables.

Variables	**Naturaleza**	**Análisis Característico**	**Árboles de Clasificación**
CLTE_Sucursal	Categórica	√	√
CLTE_Actividad_persona	Categórica	√	√
CLTE_Actividad_Empresa	Categórica	√	√
CLTE_Convenio	Categórica	√	√
ALTATC_Canal_Origen	Categórica	√	√
VERAZ_TAR_MESES_ANTIGUEDAD	Numérica	√	√
CLTE_Edad	Numérica	√	√
CLTE_provincia	Categórica	√	√
ALTATC_Limite_Compra	Numérica	√	
VERAZ_TAR_PAGOS_MIN_deflac	Numérica	√	
VERAZ_CUOTA	Numérica	√	
VERAZ_CTA_ACDO_ACTIVAS	Numérica	√	
HP_FLT_Ing_Inferido	Numérica	√	
Calificación_Concepto_Modif	Numérica	√	
CLTE_Segmento	Categórica	√	
ALTATC_alcance	Categórica	√	
CLTE_Estado_civil	Categórica	√	
VERAZ_SUCURSAL	Categórica	√	
VERAZ_PREDICTOR_NSE	Categórica	√	
VERAZ_AUTOMATICO	Categórica	√	

CALIF_Producto_Generador	Categórica	$\surd$
CALIF_Estado_Calificación_PP	Categórica	$\surd$
ALTATC_Administradora	Categórica	$\surd$
CALIF_Motivo_Descarte	Categórica	$\surd$
CALIF_Cod_Descarte	Categórica	$\surd$
VERAZ_CONSULTAS_S_P	Categórica	$\surd$

Tabla Resumen Selección de Variables

A pesar que la diferencia en cantidad de variables es notable, la relación entre variables categóricas y variables numéricas para ambos análisis se mantiene constante siendo esta aproximadamente de 3:2. Así mismo, en ámbos analisis se cuenta con las variables CLTE_Actividad_persona, CLTE_Actividad_Empresa y CLTE_Edad, las cuales son indispensables y necesarias para realizar cualquier modelo de *score* para clientes nuevos. Por lo tanto, los resultados obtenidos para la selección de variables fueron los esperados.

- **Modelo de Regresión Logística:** en relación al modelo producto de la regresión logística, se obtuvo que el modelo realizado con el conjunto de variables del análisis característico, generó un AIC de 46163, mientras que el modelo producto del conjunto de variables generadas por el árbol de clasificación arrojó un AIC de 48889. Ahora, como el criterio de información de *Akaike* es una medida de la calidad relativa de un modelo estadístico para un conjunto dado de datos, el modelo preferido es el que tiene el menor valor en el AIC. Por lo tanto, basándonos en lo descrito anteriormente, las variables que se obtienen del análisis característico predicen un mejor modelo de regresión logística.

- **Validación de los Modelos de *Score*:** con respecto a la validación, el modelo producto del análisis característico tuvo una apertura de 10 rangos en el *score*, mientras que el modelo que se realizó con las variables que resultaron predictivas para el árbol de clasificación, tuvo una apertura de 9 rangos, con lo cual se concluye que el primer modelo garantiza mejor granularidad en el estudio, sin embargo, ambos presentaron un comportamiento monótono creciente en la tasa de malos y un comportamiento monótono decreciente en los odds lo que coincide con la coherencia de la clasificación de los rangos (a mayor rango de *score*, menor tasa de malos y mayor odds; y a menor rango de *score*, mayor tasa de malos y menor odds).

  Por otra parte, en relación al KS, se obtuvo que para el primer modelo (variables con análisis característico) resultó de 40,23%, mientras que para el segundo modelo (variables con árboles

de clasificación) se tuvo un valor de 32,91%. Con lo que se concluye que a pesar que ambos modelos presentan buena distribución en el *score,* el primero tiene una mejor capacidad de discriminación de la variable respuesta, lo cual coincide con lo concluído en el ítem anterior.

INFERENCIA DE RECHAZADOS

En relación a los resultados obtenidos en la inferencia de los clientes que se encontraban rechazados por la entidad bancaria, no se hallaron mayores discrepancias entre el comportamiento inferido de los clientes utilizando el modelo proveniente del análisis característico y el modelo utilizando los árboles de clasificación. Si bien el primero de ellos fue considerado "mejor" con respecto al segundo en el apartado anterior, la diferencia entre los clientes inferidos por el modelo de árbol de clasificación no es alarmante ya que de una población de 128.814, sólo 20 registros resultaron con un comportamiento desigual.

ANÁLISIS SOBRE LA POBLACIÓN TOTAL

- **Selección de Variables:** para la selección de variables se utilizó el análisis característico y los árboles de clasificación, en donde se obtuvo que por el primer método señalado resultaron 23 variables aptas mientras que por el análisis de árboles se obtuvieron 8 variables.

Variables	Naturaleza	Análisis Característico	Árboles de Clasificación
CLTE_Segmento	Categórica	√	
ALTATC_alcance	Categórica	√	
CLTE_Convenio	Categórica	√	√
CLTE_Sucursal	Categórica	√	√
CLTE_Estado_civil	Categórica	√	
ALTATC_Canal_Origen	Categórica	√	√
CLTE_provincia	Categórica	√	√
ALTATC_Administradora	Categórica	√	
ALTATC_Sucursal	Categórica	√	
VERAZ_AUTOMATICO	Categórica	√	
CLTE_Actividad_Empresa	Categórica	√	√
VERAZ_SUCURSAL	Categórica	√	
CLTE_Actividad_persona	Categórica	√	√

CALIF_Producto_Generador	Categórica	√	
VERAZ_PREDICTOR_NSE	Categórica	√	
ALTATC_Limite_Compra	Numérica	√	
CLTE_Edad	Numérica	√	√
VERAZ_CUOTA_deflac	Numérica	√	
VERAZ_TAR_MESES_ANTIGUEDAD	Numérica	√	√
VERAZ_TAR_PAGOS_MIN_deflac	Numérica	√	
VERAZ_CTA_ACDO_ACTIVAS	Numérica	√	
HP_FLT_Ing_Inferido	Numérica	√	
Calificación_Concepto_Modif	Numérica	√	

Tabla Resumen Selección de Variables Final

A pesar que la diferencia en cantidad de variables es notable, la relación entre variables categóricas y variables numéricas para ambos análisis se mantiene constante siendo esta aproximadamente de 3:1. Así mismo, en ámbos analisis se cuenta con las variables CLTE_Acti-vidad_persona, CLTE_Actividad_Empresa y CLTE_Edad, las cuales son indispensables y necesarias para realizar cualquier modelo de *score* para clientes nuevos. Por lo tanto, los resultados obtenidos para la selección de variables fueron los esperados.

- **Modelo de Regresión Logística:** en relación al modelo producto de la regresión logística, se obtuvo que el modelo realizado con el conjunto de variables del análisis característico, generó un AIC de 48198, mientras que el modelo producto del conjunto de variables generadas por el árbol de clasificación arrojó un AIC de 49611. Ahora, como el criterio de información de *Akaike* es una medida de la calidad relativa de un modelo estadístico para un conjunto dado de datos, el modelo preferido es el que tiene el menor valor en el AIC. Por lo tanto, basándonos en lo descrito anteriormente, las variables que se obtienen del análisis característico predicen un mejor modelo de regresión logística.

- **Validación de los Modelos de *Score*:** con respecto a la validación, ambos modelos tuvieron una apertura de 10 rangos en el *score* lo cual garantiza granularidad en el estudio, así mismo, presentaron un comportamiento monótono creciente en la tasa de malos y un comportamiento monótono decreciente en los odds lo que coincide con la coherencia de la clasificación de los rangos (a mayor rango de *score*, menor tasa de malos y mayor odds; y a menor rango de *score*, mayor tasa de malos y menor odds).

Sin embargo, en relación al KS, se obtuvo que para el primer modelo (variables con análisis característico) resultó de 37,68%, mientras que para el segundo modelo (variables con árboles de clasificación) se tuvo un valor de 35,1%. Con lo que se concluye que a pesar que ambos modelos presentan buena distribución en el *score,* el primero tiene una mejor capacidad de discriminación de la variable respuesta, lo cual coincide con lo concluído en el ítem anterior.

ANÁLISIS DEL *SCORE* FINAL

Para definir el umbral de *score* donde la entidad bancaria ubicará a los clientes según su comportamiento, se halla en las tablas de validación total el rango de *score* en donde el valor del KS sea el máximo, es decir, si se observa las tablas total de los modelos, se concluye que un cliente se considerará bueno si su *score* es mayor o igual a 503 (Rango 3) para el caso del modelo con análisis característico y mayor o igual a 507 (Rango 3) para el modelo realizado con la metodología de árboles de clasificación; por el contrario el cliente será malo si su *score* es menor que estos valores.

En general, al tener que ambos modelos presentan el mismo umbral de puntuaciones para definir el comportamiento de sus clientes, se considerará únicamente los argumentos previos para concluir que para los datos suministrados por la entidad bancaria argentina, la selección de variables más óptima resultó del análisis característico ya que el conjunto de variables que resultaron predictivas con esta metodología, generaron un mejor modelo de *score* y una discriminación de clientes buenos y malos más alta con respecto al modelo realizado con las variables producto del árbol de clasificación.

En relación a las recomendaciones para trabajos futuros se sugiere:

- Usar los modelos desarrollados con datos actuales de la entidad bancaria para verificar si aún presenta validez su discriminación.
- Generar un árbol de clasificación con más variables predictivas, para conocer si mayor cantidad de variables conduce a un mejor modelo estadístico.
- Verificar los resultados obtenidos en la tabla de validación ya que de lo contrario se pueden incurrir en suposiciones erróneas en cuanto a la capacidad de discriminación de un modelo.
- Utilizar alguna otra metodología para la clasificación de variables y comparar con las dos expuestas en este trabajo, ya que se estaría proporcionando más alternativas válidas que facilitan la regresión logística al momento de realizar modelos de *scoring*.

REFERENCIAS

[1] Siddiqi, Naeem, Credit Risk Scorecards, **Developing and Implementing Intelligent Credit Scoring**. John Wiley & Sons, Nueva Jersey, 2006.

[2] Kleinbaum, Kupper, Nizam, Rosenberg, **Applied Regression Analysis and Other Multivariable Methods**. Cencage Learning, Boston, 2014.

[3] Marin JM, **Análisis de Cluster y Arboles de Clasificación**. Disponible en Internet: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf

[4] Hothorn T, Hornik K, Zeileis A, **ctree: Conditional Inference Trees.** Alemania, 2006. Disponible en Internet: https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf