ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS
DEPARTAMENTO DE POSTGRADO

PROYECTO DE TITULACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

“MAGÍSTER EN ESTADÍSTICA CON MENCION EN GESTIÓN DE LA CALIDAD Y PRODUCTIVIDAD”

TEMA:

“Medición del riesgo y predicción del incumplimiento de pago de clientes en una empresa de Guayaquil para el año 2015 utilizando técnicas de minería de datos”

AUTOR:

ROSA NATHALÍ MANCERO CASTRO

Guayaquil - Ecuador

2018

DEDICATORIA

A Dios, mis padres, familia, amigos y a mis jefaturas que me han brindado su apoyo incondicional ofreciéndome sus palabras de aliento y todo su afecto para poder culminar una de las etapas más importantes de mi vida.

Rosa Nathalí Mancero Castro.

AGRADECIMIENTO

A Dios, por permitirme culminar una nueva etapa de mi vida profesional, a mis padres José Mancero Guevara, María Barrera Mendoza e Isabel Castro Macías por brindarme su apoyo incondicional para seguir adelante. A mis hermanos, amigos, jefaturas y al amor de mi vida Juan quienes han sido pilares fundamentales en este proceso.

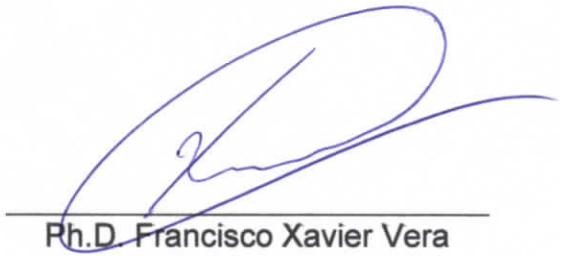
Rosa Nathalí Mancero Castro

DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Graduación, me corresponde exclusivamente; el patrimonio intelectual del mismo, corresponde exclusivamente a la **Facultad de Ciencias Naturales y Matemáticas, Departamento de Postgrado** de la Escuela Superior Politécnica del Litoral.

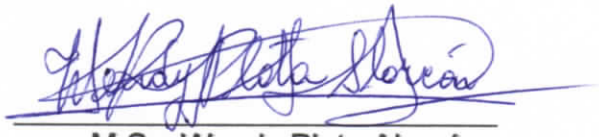
Rosa Nathalí Mancero Castro

TRIBUNAL DE GRADUACIÓN



Ph.D. Francisco Xavier Vera
Presidente

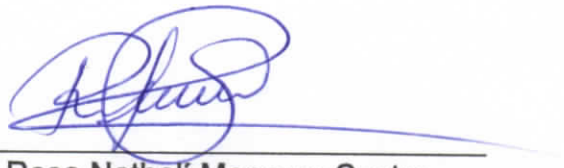
M.Sc. Brenda Denisse Cobeña
Director



M.Sc. Wendy Plata Alarcón
Vocal Principal 1

Msc. Francisco Moreira Villegas
Vocal Principal 2

AUTOR DEL PROYECTO



Rosa Nathalí Mancero Castro

ÍNDICE GENERAL

ÍNDICE DE TABLAS

INDICE DE FIGURAS

PRESENTACIÓN

La importancia para la empresa de estudio ubicada dentro del top 10 de las empresas comerciales más importantes de la ciudad de Guayaquil de poseer modelos predictivos para su gestión del riesgo de crédito es vital en la era actual, debido a que la competitividad es cada vez mayor entre las empresas en base a la información y el conocimiento, lo cual le ha conllevado a la necesidad de analizar grandes volúmenes de datos y adquirir nuevas herramientas de data mining o minería de datos, con el objetivo de explotar los beneficios en base a la toma de decisiones para así tener ventajas mercadológicas frente a los competidores.

Más del 60% del portafolio de la empresa de nuestro estudio corresponde a ventas a crédito y dado el giro del negocio existe la necesidad de implementar modelos predictivos que permitan disminuir el riesgo al momento de dar el crédito, para así minimizar los tiempos de respuesta, reducir costos y a su vez maximizar la rentabilidad de la entidad mediante la corrección de la pérdida esperada de la institución.

Para ello, el presente trabajo pretende justificar con robustez estadística la probabilidad de que un cliente pueda caer en incumplimiento y así evitar pérdidas de liquidez para la compañía mediante el uso de varias técnicas estadísticas como son: Regresión Logística, Arboles de decisión, Bagging y Bosques Aleatorios; y de esta forma identificar las variables que infieren en el comportamiento de pago de sus clientes y así definir perfiles o políticas claramente establecidas sobre los clientes catalogados como buenos o malos pagadores.

Es relevante indicar que los modelos predictivos presentados en este análisis evalúan a los clientes al momento del otorgamiento y el comportamiento del crédito con el objetivo de analizar el desempeño inicial hasta el cumplimiento esperado final de la obligación crediticia que poseen los clientes con la empresa.

CAPÍTULO 1

INTRODUCCIÓN

1.1. ANTECEDENTES

En los años 40 Thomas, Crook & Edelman (2002), establecían que las metodologías de discriminación y clasificación podían ser aplicadas en la evaluación del crédito como cualquier área de estudio. Luego en los años 50 se comenzaron a utilizar modelos estadísticos en las decisiones de préstamos. Bill Fair y Earl Isaac instituyen para ese entonces su compañía dedicada a colaborar en las actividades de empresas financieras importantes y de ventas al por menor.

Ya, en los años 1960's inicia la fase en la que se desarrolla la comercialización de las tarjetas de crédito con lo cual las entidades bancarias ven una gran oportunidad de comenzar a manejar dichos modelos debido al incremento de solicitudes que se debían ejecutar constantemente. (Thomas, A survey of credit and behavioural scoring forecasting risk of lending to consumers, 2000).

Finalmente en los años 80, dado al gran éxito presentado en las tarjetas de crédito, se empezaron a aplicar los modelos sobre otros servicios como préstamos personales, artículos para el hogar, anticipos a medianas empresas entre otros. En este periodo se usaban modelos computarizados debido al incremento de la oferta y la demanda, ya que existía la necesidad de apoyarse en la automatización de los modelos tradicionales tales como regresión logística y la programación lineal.

En la actualidad el análisis de riesgo de crédito es uno de los temas más trascendentales en el campo de la gestión financiera y comercial en donde la capacidad para clasificar a los clientes según su comportamiento crediticio es

crucial. Para este fin es necesaria la evaluación de modelos cuantitativos confiables que predigan los incumplimientos con el menor error posible para que la alta gerencia y ejecutivos puedan tomar medidas preventivas.

La competencia agresiva del mercado, la liquidez y la fidelización de los clientes son algunos de los problemas actuales para la empresa de este proyecto. Por ello, se busca construir un modelo que permita automatizar las aprobaciones de crédito en el menor tiempo posible. El mejor modelo de clasificación permitirá maximizar la concreción de las ventas y reducir el riesgo de colocar los portafolios de sus productos en segmentos de clientes con menores posibilidades de pago a nivel país.

En Ecuador existen consultorías o empresas que se dedican a ofrecer soluciones mediante ajustes de modelos matemáticos tales como: Logit, Z Altman, árboles de decisión y regresión múltiple para la predicción de clientes impagos a empresas comerciales y financieras con costos excesivamente elevados. Por tanto, existe la necesidad de optimizar recursos y elaborar modelos internos para así incurrir en gastos innecesarios utilizando modelos estadísticos de vanguardia a menor costo.

Estadísticos y científicos de datos crean constantemente nuevas herramientas o metodologías analíticas que ayudan a mejorar la predicción. Estos modelos de vanguardia pueden contribuir con el descubrimiento de nuevos patrones en los perfiles de pago de los clientes. Algunas de las técnicas que pueden usarse para la evaluación del riesgo de crédito son: Regresión Logística, Árboles de clasificación, Bagging, Bosques Aleatorios, Máquinas de soporte vectorial entre otros.

Estas técnicas pueden complementar la técnica tradicional de las ‘’cinco C’’ que se basa en las siguientes consideraciones:

1. **<u>Condiciones Económicas:</u>** Se refiere a las condiciones económicas o a factores exógenos que son determinantes al momento del otorgamiento del crédito.

2. **Capacidad de Pago:** Se refiere al ingreso, capacidad de endeudamiento o fuente de ingresos.

3. **Capital:** Son los bienes que el cliente posee.

4. **Colateral:** Son las garantías que avalúan el préstamo.

5. **Carácter:** Se refiere a la honorabilidad y solvencia moral del cliente en donde a través del historial de las referencias del buró se evidencia tanto el sector financiero regulado (bancos), no regulado (cooperativas) y casas comerciales, si el cliente cumple o no sus obligaciones crediticias.

La técnica tradicional de las cinco C, es aplicada actualmente por las jefaturas y analistas de crédito de la compañía en donde realizan evaluaciones subjetivas, no solo revisando la información de referencias crediticias externas sino también analizando el comportamiento actual e histórico de pago de los clientes internamente con la empresa. La desventaja de esta metodología es que se puede aplicar solo para poblaciones pequeñas o para un tipo de crédito en específico.

En el segundo capítulo de este trabajo explicaremos sobre los fundamentos matemáticos de las cuatro técnicas que se usarán en este estudio: Regresión logística, Árboles de decisión, Bosques aleatorios y Bagging. Describiremos también cómo se implementaron estas técnicas en el software estadístico.

En el capítulo tres, se analiza a la población objetivo de este estudio usando las técnicas estadísticas definidas en el capítulo dos. Se describirá el esquema de desarrollo y se obtendrán las tablas de confusión, indicadores de precisión, error, gráficas y las variables significativas o relevantes según cada modelo. Adicionalmente se presenta la aplicación del modelo credit scoring que usa regresión logística, junto con las definiciones de medidas de calidad (KS, ROC y GINI.), puntos de corte, exclusiones, variables dependiente, independiente, ecuación de regresión, tablas performance y scorecards.

Para cada modelo estadístico se presentará el siguiente esquema de variables:

Modelo I o Modelo de Originación: Evalúa el desempeño de pago sobre las variables demográficas del cliente y hace referencia a las variables que se ingresan en la solicitud al momento del otorgamiento del crédito.

Modelo II o Modelo de Comportamiento: Evalúa el desempeño de pago de los clientes en base a variables demográficas y de comportamiento financiero. El uso de ambos conjuntos de variables puede ayudar a mejorar la predicción de si el cliente caerá o no en morosidad.

El capítulo cuarto, muestra las conclusiones y recomendaciones en base a los perfiles de clientes morosos y la mejor metodología de predicción para nuestra data.

Finalmente, se presenta una solución a la problemática actual del otorgamiento de crédito de la entidad asumiendo un nivel de riesgo esperado.

1.2. PLANTEAMIENTO DEL PROBLEMA

El desempleo, las enfermedades, sobreendeudamientos, la muerte, daños en los productos, cambios de domicilio, y robos son algunos de los factores por los cuales los clientes pierden la voluntad de cumplir con sus obligaciones crediticias a largo plazo con la entidad. Estas situaciones conllevan a un incremento del riesgo de afectación de liquidez y rentabilidad de la compañía. Por otro lado, la presión en los indicadores de cumplimiento en el área comercial y la necesidad de ganar comisiones, comúnmente conducen de manera equivocada a flexibilizar las políticas internas de crédito que como resultado genera sobreendeudamiento en los productos e incumplimientos futuros en las promesas de pago de sus clientes.

Ante la problemática actual, el departamento de crédito de la compañía definió como prescindible el uso de modelos de scoring genéricos los cuales son elaborados mediante test psicométricos que evalúan la relación de la personalidad del individuo con el comportamiento de pago o modelos en base al comportamiento real de pago en el SCE (Sistema Crediticio Ecuatoriano), es decir en bancos, casas comerciales, cooperativas de ahorro entre otros.

La ventaja de los scores genéricos es que ayudan a minimizar el tiempo de respuesta en la aprobación del crédito, pero al no ser un modelo a la medida su error de predicción es más alto, dado que no consideran las variables demográficas y de comportamiento de pago de los clientes de la entidad. En este caso, para los clientes de este análisis la empresa decidió aplicar el modelo genérico de scoring psicométrico en donde se evidenció que la tasa de error de predicción bordeaba alrededor del 50% aproximadamente.

Adicionalmente, los modelos a la medida elaborados por consultorías son costosos. Por tanto en este estudio se consideró la construcción de modelos predictivos con datos internos con el objetivo de obtener un menor error predictivo que el actual. Esto es primordial para la empresa para que exista un punto de equilibrio entre su margen de ventas y la gestión de cobranza sin afectar la utilidad bruta neta de la empresa.

1.3 MOTIVACIÓN

Los motivos para realizar el presente estudio de morosidad se mencionan a continuación:

- Hoy en día para el segmento de clientes no bancarizados o sin información crediticia se deja de facturar aproximadamente el 10% respecto al total de su cartera generada por ende se pretende maximizar la colocación una vez generado el modelo.
- Se prevé reducir la tasa de incobrabilidad dentro de la institución en un 15% estimado respecto a la incobrabilidad actual mediante la implementación del mejor modelo de decisión obtenido.
- Mejorar la concreción de las ventas, otorgándoles crédito directo inmediato mediante el correcto perfilamiento de clientes.
- Automatizar el modelo para agilizar las aprobaciones de crédito dentro de la entidad.
- Establecer un modelo a la medida de acuerdo al apetito de riesgo de la entidad logrando que la rentabilidad y la perdida esperada no pierdan su equilibrio.

1.4 OBJETIVOS PLANTEADOS

Proveer de información predictiva a los Directivos de la Compañía para la toma de decisiones, como un mecanismo en la optimización de colocación de créditos y reducción de índices de morosidad, mediante la obtención de una mejor clasificación de clientes según su desempeño de pago.

- Definir el mejor modelo estadístico de otorgamiento in house que permita identificar los posibles clientes con riesgo de morosidad.
- Determinar la probabilidad de morosidad de los clientes y de esta forma tomar acciones preventivas.

1.3. ALCANCE

El alcance del presente proyecto está dirigido a clientes sin información crediticia en el sector financiero o comercial y que hayan adquirido por primera vez crédito con la entidad.
El periodo de análisis para la implementación de las diferentes metodologías estadísticas se encuentra comprendida entre enero y diciembre 2015 los cuales hacen referencia a variables demográficas y de comportamiento de pago de sus clientes. Cabe destacar que la data será tratada con estricta confidencialidad.

1.4. REVISIÓN BIBLIOGRÁFICA

Para la evaluación del riesgo de crédito se han implementado diversas técnicas estadísticas de regresión y clasificación. Algunas de estas técnicas tienen más bien un origen computacional, tales como: redes neuronales o máquinas de soporte vectorial (SVM). Cada metodología tiene sus ventajas y limitaciones y sus resultados dependen de la estructura de datos analizados, por lo que no se pueden generalizar en otros problemas.

SVM, es una técnica creada por Vapnik (1995) y se ha convertido en una herramienta de gran importancia para solucionar problemas de clasificación y regresión. Algunos investigadores como Bellotti, Crook (2009) y Yu et al. (2010), han utilizado SVM para solucionar el problema de riesgo de crédito (Gutierrez & Melo, 2011). Existen aplicaciones más recientes de SVM en el contexto de riesgo de crédito por ejemplo: Harris (2015) introduce el uso de clústers dentro de la metodología de máquinas de soporte vectorial (CSVM) para la colocación de tarjetas de crédito. Por su parte Zhang et al. (2015), hace una comparación entre el modelo máquinas de soporte vectorial (SVM) y la técnica de Redes Neuronales para la cadena de suministro de finanzas (SCF), en donde se concluye que el modelo más óptimo para evaluar el riesgo de crédito de sus clientes pertenecientes al sector PYMES (Pequeñas y medianas empresas) es la de máquinas de soporte vectorial.

Los autores Yu et al. (2008) Presentan una revisión extensa de la literatura entre los años 1970 al 2017 de aproximadamente 600 trabajos y libros especializados en scoring y evaluación de riesgo de crédito entre los cuales se destacan: Evaluación del riesgo de crédito mediante SVM con búsqueda directa para la selección de parámetros y un enfoque SVM difuso de mínimos cuadrados para la evaluación del riesgo de crédito. Debido a que cada técnica posee una estructura diferente, es difícil señalar cual es la idónea, más sin embargo de las investigaciones y trabajos realizados sobre scoring se concluye que SVM posee mayor flexibilidad que las metodologías de regresión y discriminación e inclusive

(Li, Wei, & Hao, 2013), proponen a Bagging mejorado en base a atributos seleccionados por peso (WSAB), para evaluar el riesgo de crédito. En esta contribución contrastan los pesos para cada atributo en base a las metodologías de máquinas de vectores de soporte lineal (LSVM) y el análisis de componentes principales (PCA), en donde, en base a los resultados experimentales obtenidos en dos bases de datos que detallan el historial crediticio de clientes en una entidad, concluyen que el método propuesto, WSAB, es sobresaliente tanto en precisión de predicción como en estabilidad, en comparación con los métodos análogos descritos.

Regresión Logística, por su lado, estima la probabilidad de ocurrencia de un suceso a partir de un grupo de variables de explicación. Es una de las metodologías más utilizadas por su fácil interpretación, aceptación y por la contribución de sus resultados en el riesgo de crédito. Regresión logística ofrece adicionalmente puntajes o scores de crédito para la aprobación inmediata al momento del otorgamiento. En la literatura podemos encontrar algunos trabajos que usan regresión logística, en este contexto por ejemplo, Agbemava et al. (2016), identifica los factores de riesgo asociados a los clientes impagos en una institución financiera de Ghana. Entre ellos se destacan el estado civil, tipo de garantía, tipo de préstamo y el plazo del crédito. Por su parte Abid et al. (2016), señala que existen 3 características que definen la probabilidad de que los prestarios en una institución bancaria incumplan con el pago: el monto del préstamo, deudas pendientes y profesión.

(Vera, Camanho, & Borgues, 2017) Comparten sus experiencias en el desarrollo, la implementación y la evaluación de sus decisiones y estrategias comerciales, proponiendo un modelo de respuesta de minería de datos basado en bosques aleatorios para la definición de clientes objetivos para campañas bancarias. En este estudio también se midió el rendimiento de varias metodologías de remuestreo y otros modelos de minería de datos. Este estudio concluye que bosques aleatorios posee un mejor rendimiento de predicción en sus datos.

(Chen, Yeong, & Zone, 2014) , compararon las metodologías de regresión logística, máquina de soporte vectorial (SVM) y árboles de decisión para predecir estados financieros fraudulentos de una empresa de estudio, en el periodo entre 1998 y 2012. Los autores concluyen que la metodología de árboles de decisión obtuvo la mejor tasa de predicción y clasificación con un 85.71%.

Finalmente, en base a las conclusiones y hallazgos de los expertos para cada una de las metodologías de minería de datos, lo que se busca dentro de este contexto académico, es determinar el mejor modelo que permita a la empresa optimizar la aprobación de los créditos de forma automática y que prediga con el menor error posible la capacidad de pago de un cliente.

CAPÍTULO 2

METODOLOGÍA

2.1 INTRODUCCIÓN

En este capítulo, se define a la población objetivo, las variables de estudio, la definición de la variable dependiente e independiente, los tipos de modelos de riesgo a implementar y la connotación científica sobre las metodologías estadísticas que se llevaran a cabo en el propósito de este estudio.

Los procesos generales contemplados para la elaboración de modelos analíticos de este estudio se basaron en los siguientes procedimientos generales: (Ver Figura 1):

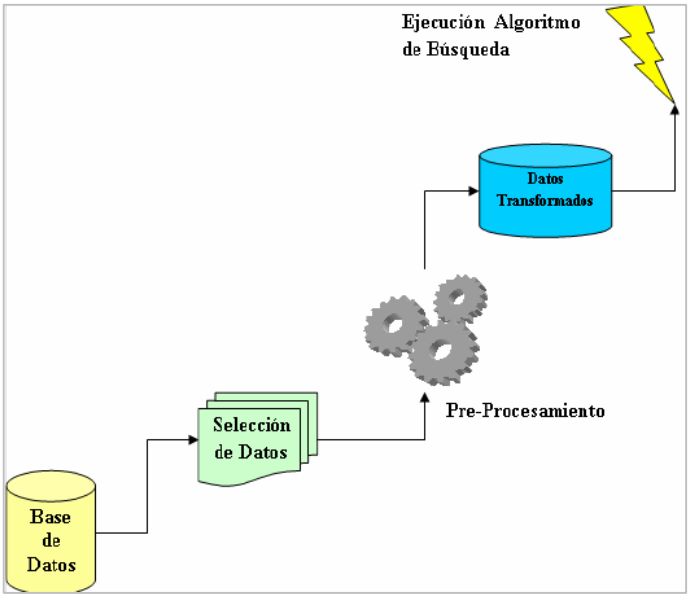
- **Selección de datos:** Se define la selección de las variables más relevantes de estudio en base a la población objetivo y al tipo de modelo.

- **Pre-Procesamiento:** En esta etapa se procede con la limpieza de los datos y el análisis de la muestra en donde se evalúa la aleatoriedad y representatividad de los datos referente a la población.

- **Datos Transformados:** No, existió la necesidad de realizar el proceso de normalización ($\frac{(X_i-\mu)}{\sigma}$.)de variables debido que no existían rangos muy elevados.

- **Ejecución Algoritmos de Búsqueda**: Finalmente se ejecuta la Programación en Rstudio para cada una de las metodologías de minería de datos para la implementación de este análisis tales como: Regresión Logística, Arboles de decisión, Bosques aleatorios y Bagging. (Hastie, Tibshirani, & Friedman, 2009)

Figura 1: Proceso de predicción del riesgo de crédito



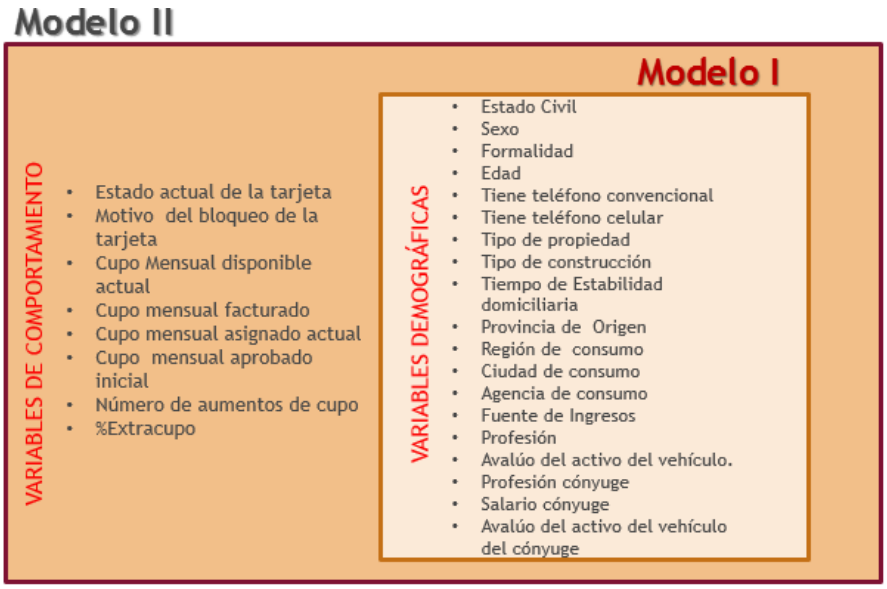
Fuente: Soto, 2016

Para la predicción de los clientes con buen y mal comportamiento de pago, se consideró una muestra representativa de clientes no bancarizados de la entidad con año de aprobación y consumo de crédito 2015, en edades mayores a 18 años con estabilidad domiciliaria e ingresos económicos demostrables por fuente laboral o negocio.

Las variables relevantes de estudio para el respectivo modelamiento se clasifican en 2 secciones:

Variables de Identificación: Se consideran como variables independientes y contienen la información demográfica o propia de los clientes tales como: Sexo, Edad, Fecha de Nacimiento, Formalidad, etc. y adicional está conformada por variables de comportamiento de pago como por ejemplo: Estado actual de la tarjeta, Motivo del bloqueo de la tarjeta, Cupo mensual disponible actual, etc. (Ver. Figura 2). En la sección del capítulo 3, se detallan las variables que se utilizarán para los modelo de originación y comportamiento junto con su respectiva codificación.

Figura 2: Variables de Identificación para los Modelos I y II



Fuente: Elaboración Propia

Variables de Desempeño: Permitirá definir la variable dependiente, es decir, Si el cliente debe ser catalogado como moroso $(Y = 1)$ o no $(Y = 0)$. Estas variables reflejan el cumplimiento de pago de los clientes en una ventana o periodo de análisis de 12 meses.

Dentro del set de la base de datos constan por ejemplo: Monto de crédito desembolsado y días de mora desde el periodo 1 hasta el 12. (Ver Figura 3):

En la sección del capítulo 3, se muestra el procedimiento del cálculo para la obtención de la variable dependiente mediante las metodologías de matrices de transición y roll rate. La definición de la variable dependiente, aplica para los modelos I y II:

Modelo I o Modelo de Originación: Evalúa el desempeño de pago sobre las variables demográficas del cliente y hace referencia a las variables que se ingresan en la solicitud al momento del otorgamiento del crédito.

Modelo II o Modelo de Comportamiento: Evalúa el desempeño de pago de los clientes en base a variables demográficas y de comportamiento financiero. En conjunto ambos sets de variables ayudarán a encontrar la probabilidad más exacta, de si, el cliente caerá en mora o no.

Figura 3: Variables de Comportamiento para la definición de la Variable dependiente: ''Morosidad''

Nombre	Descripción
Variables en la ventana de desempeño a nivel de OPERACIÓN	
Id. Cliente	Id que identifique al cliente a cambio del número de cédula
Num_ope	Id que identifique la operación del cliente
Pto_Obs	Fecha de evaluación del sujeto
Fecha_aprobación	Fecha de aprobación del crédito
fecha_otorgamiento	fecha de concesión del crédito
plazo	plazo en el cual se concedió el crédito
Calificación Compra	Calificación al Momento de la compra
Calificación Actual	Calficación Interna Actual del cliente por comportamiento de pago
monto_credito	monto del crédito
Cuota uno Impaga	Falta de pago en la primera compra
saldo_financiar	saldo a financiar
sal_deuda_n1	Suma del valor xvencer + ndi + vencido+ demanda + castigo al mes siguiente del pto de obs
sal_xvencer_n1	Valor por vencer al mes siguiente del pto de obs
sal_vencido_n1	Valor vencido al mes siguiente del pto de obs
num_dven_n1	numero de días vencidos al mes siguiente del pto de obs
sal_deuda_n2	Suma del valor xvencer + ndi + vencido + demanda + castigo al segundo mes siguiente del pto de obs
sal_xvencer_n2	Valor por vencer al segundo mes siguiente del pto de obs
sal_vencido_n2	Valor vencido al segundo mes siguiente del pto de obs
num_dven_n2	numero de días vencidos al segundo mes siguiente del pto de obs
sal_deuda_n3	Suma del valor xvencer + ndi+ vencido + demanda + castigo al tercer mes siguiente del pto de obs
sal_xvencer_n3	Valor por vencer al tercer mes siguiente del pto de obs
sal_vencido_n3	Valor vencido al tercer mes siguiente del pto de obs
num_dven_n3	numero de días vencidos al tercer mes siguiente del pto de obs
sal_deuda_n4	Suma del valor xvencer + ndi + vencido + demanda + castigo al quarto mes siguiente del pto de obs
sal_xvencer_n4	Valor por vencer al quarto mes siguiente del pto de obs
sal_vencido_n4	Valor vencido al quarto mes siguiente del pto de obs
num_dven_n4	numero de días vencidos al quarto mes siguiente del pto de obs
sal_deuda_n5	Suma del valor xvencer + ndi + vencido + demanda + castigo al quinto mes siguiente del pto de obs
sal_xvencer_n5	Valor por vencer al quinto mes siguiente del pto de obs
sal_vencido_n5	Valor vencido al quinto mes siguiente del pto de obs
num_dven_n5	numero de días vencidos al quinto mes siguiente del pto de obs
sal_deuda_n6	Suma del valor xvencer + ndi + vencido + demanda + castigo al sexto mes siguiente del pto de obs
sal_xvencer_n6	Valor por vencer al sexto mes siguiente del pto de obs
sal_vencido_n6	Valor vencido al sexto mes siguiente del pto de obs
num_dven_n6	numero de días vencidos al sexto mes siguiente del pto de obs
sal_deuda_n7	Suma del valor xvencer + ndi + vencido + demanda + castigo al septimo mes siguiente del pto de obs
sal_xvencer_n7	Valor por vencer al septimo mes siguiente del pto de obs
sal_vencido_n7	Valor vencido al septimo mes siguiente del pto de obs
num_dven_n7	numero de días vencidos al septimo mes siguiente del pto de obs
sal_deuda_n8	Suma del valor xvencer + ndi + vencido + demanda + castigo al octavo mes siguiente del pto de obs
sal_xvencer_n8	Valor por vencer al octavo mes siguiente del pto de obs
sal_vencido_n8	Valor vencido al octavo mes siguiente del pto de obs
num_dven_n8	numero de días vencidos al octavo mes siguiente del pto de obs
sal_deuda_n9	Suma del valor xvencer + ndi + vencido + demanda + castigo al noveno mes siguiente del pto de obs
sal_xvencer_n9	Valor por vencer al noveno mes siguiente del pto de obs
sal_vencido_n9	Valor vencido al noveno mes siguiente del pto de obs
num_dven_n9	numero de días vencidos al noveno mes siguiente del pto de obs
sal_deuda_n10	Suma del valor xvencer + ndi + vencido + demanda + castigo al decimo mes siguiente del pto de obs
sal_xvencer_n10	Valor por vencer al decimo mes siguiente del pto de obs
sal_vencido_n10	Valor vencido al decimo mes siguiente del pto de obs
num_dven_n10	numero de días vencidos al decimo mes siguiente del pto de obs
sal_deuda_n11	Suma del valor xvencer + ndi + vencido + demanda + castigo al decimo primero mes siguiente del pto de obs
sal_xvencer_n11	Valor por vencer al decimo primero mes siguiente del pto de obs
sal_vencido_n11	Valor vencido al decimo primero mes siguiente del pto de obs
num_dven_n11	numero de días vencidos al decimo primero mes siguiente del pto de obs
sal_deuda_n12	Suma del valor xvencer + ndi + vencido + demanda + castigo al decimo segundo mes siguiente del pto de obs
sal_xvencer_n12	Valor por vencer al decimo segundo mes siguiente del pto de obs
sal_vencido_n12	Valor vencido al decimo segundo mes siguiente del pto de obs
num_dven_n12	numero de días vencidos al decimo segundo mes siguiente del pto de obs

Fuente: Elaboración Propia

Para la aplicación de los modelos I y II, se implementarán metodologías estadísticas que puedan explicar la variable a analizar. En este caso como las variables de estudio son de tipo nominal, ordinal y de escala, se recomienda utilizar los siguientes modelos de minería de datos: Regresión Logística (RGL), Árboles de decisión, Bagging y Bosques Aleatorios.

Se recuerda que es importante recordar que el tipo de variable [1] influye en la técnica a considerar.

A continuación presentamos un ejemplo de algunas metodologías estadísticas aplicadas de acuerdo al tipo de variable en función de la variable dependiente e Independiente del modelo. Nótese que los modelos en rojo son las técnicas que aplicaremos en este estudio.

Figura 4: Metodologías Estadísticas Aplicadas de Acuerdo al Tipo de Variable

		Variables Independientes (*X*)	
		Variables Cualitativas (nominales u ordinales)	Variables Cuantitativas (De escala)
Variable Dependiente (*Y*)	Variables Cualitativas (nominales u ordinales)	Análisis Discriminante, Regresión Logística, etc.	Análisis Discriminante,
	Variables Cuantitativas (de escala)	ANOVA, MANOVA, MLG, Árboles de decisión , Bosques Aleatorios, Bagging	Regresión Lineal (Simple/ Múltiple), Árboles de decisión , Árboles de decisión , Bosques Aleatorios, Bagging, ANCOVA, MLG

Fuente: Elaboración Propia

La finalidad de contrastar las cuatro metodologías estadísticas, son con el objetivo de contrastar el poder de discriminación de cada modelo en base a la tasas de precisión y error para así obtener el modelo que mejor diferencie de clientes buenos y malos pagadores.

Para mejorar la optimización de los modelos de ser necesario, se realizará la transformación a las variables (Aplicación de logaritmos, raíces cuadradas, etc.) y adicionalmente para el modelo de regresión logística se presentaran otras

[1] *ANOVA (Análisis de Varianzas): Permite determinar la existencia de diferencias significativas entre más de dos grupos /tratamientos.*
ANCOVA (Análisis de Covarianzas): Fusión de la ANOVA con Regresión Lineal Múltiple
MANOVA (Análisis Multivariante de Varianzas): Extensión de la ANOVA pero para más de 1 variable dependiente
MLG (Modelo Lineal Generalizado): Generalización flexible de la Regresión Lineal Ordinaria.

medidas de calidad adicionales a ROC como son: KS y GINI junto con las tablas performance del modelo.

Finalmente para el mejor entendimiento de cada una de las metodologías se procederá con la explicación matemática de cada modelo y con la definición de cadenas de Markov para la obtención de la variable dependiente.

2.2 MODELO DE REGRESIÓN LOGÍSTICA (RGL)

Regresión logística (RGL), es un modelo lineal generalizado (GLM) que permite predecir la probabilidad de que ocurra un evento en función de varios factores. El análisis de regresión lineal múltiple tiene la misma estrategia que el análisis de regresión logística, debido a que solo se diferencian en la variable dependiente ya que en regresión lineal múltiple, la variable dependiente es cuantitativa.

CONCEPTOS PREVIOS

El modelo consta de dos tipos de variables:
Variable dependiente (Y) , y variables independientes los cuales denotaremos como $(x_1, x_2 \ldots x_q)$. La variable dependiente o también denominada "de respuesta" es de tipo discreta dicotómica (generalmente se codifica para que tome valores (Y=*1 y* Y =*0*), mientras que las variables de explicación o variables independientes pueden ser de carácter cuantitativo o cualitativo.
La ecuación del modelo no es lineal; si bien, solo por transformación logarítmica se puede presentar como una función lineal. Las variables de respuesta pueden tomar únicamente dos valores: 1, "presencia" con probabilidad p; y ,0 "ausencia" con probabilidad $(1-p)$.

2.2.1 DEFINICIÓN

(Everitt, 1998), colaboró con la siguiente definición para la distribución logística:

"En el límite, la función distribución de probabilidad, cuando n tiende al infinito, del promedio de los valores más grandes a los valores más pequeños de muestra de tamaño n vienen de una distribución exponencial."

Una variable aleatoria x se dice que tiene una Distribución Logística con parámetro θ, si y solo si:

$$f(x) = \frac{e^{-(x-\boldsymbol{\theta})}}{(1+e^{-(x-\boldsymbol{\theta})})^2}; \qquad con\ Soporte\ S = R;\ \theta \in R$$

A continuación se muestra la ecuación cuando θ es cero denominándose simplemente "Distribución Logística" (Gaudencio, 2008), cuya distribución está dada por:

$$f(x) = \frac{e^{-z}}{(1+e^{-z})^2};\ \ donde\ \ \ S = R$$

Y su Distribución Acumulada se denota como:

$$F(x) = \frac{1}{1+e^{-z}};\ \ donde\ \ z \in R$$

"La Regresión Logística está basado en la siguiente transformación logística,

$$Logit\ (p) = \frac{p}{1-p}$$

Donde;

$$p = P(Y = 1)$$

$$(1-p) = P(Y = 0)$$

La función de Regresión Logística es la transformación de P.

Donde;

$$z = Logit\ (P) = ln\left(\frac{p}{1-p}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_1 + \boldsymbol{\beta_2} x_2 + \cdots + \boldsymbol{\beta_q} x_q$$

$\boldsymbol{\beta_0}$, Es la constante de la ecuación, y los $\boldsymbol{\beta}_{\mathrm{q}}$ son los coeficientes de las variables predictoras x_q.La transformación logística es usada para evitar que las probabilidades ajustadas se encuentren fuera del rango [0,1]. (Abelraham, 2010).

Para la estimación de los coeficientes del modelo logístico se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que maximicen la probabilidad de obtener los valores de la variable dependiente Y proporcionados con los datos de la muestra.

Para la determinación de los estimadores de máxima verosimilitud se recurre a métodos iterativos, como el de Newton-Raphson, dado que el cálculo es complejo por no existir soluciones explícitas para los estimadores de $\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \ldots . \boldsymbol{\beta_q}$**;** normalmente hay que recurrir al uso de rutinas de programación basado en el texto de (Hosmer & Lemeshow).

2.2.2 MÉTODOS DE SELECCIÓN DE VARIABLES REGRESORAS

En la disposición de un conjunto grande de posibles variables regresoras, es importante conocer el número de variables que deben ingresar o no dentro del modelo, dado que la varianza del modelo $\left(\frac{RSS}{n-p-1}\right)$, influye directamente en el número de variables, lo que implica que a mayor número de variables regresoras menor es la capacidad predictiva del modelo por presencia de variabilidad. (Vilar, 2006).

En el contraste de la prueba de hipótesis para la selección de variables regresoras se establece lo siguiente:

- $H_0: \boldsymbol{\beta_1} = \cdots \boldsymbol{\beta_i}$; $(Ningun\ Xp, es\ útil\ para\ predecir\ Y)$
- $H_1: \boldsymbol{\beta_i} \neq 0; (Al\ menos\ un\ Xp\ es\ útil\ para\ predecir\ Y\ )$

Para ello, el estadístico de prueba resultante es:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p.n-p-1}$$

Donde, n es el tamaño de la muestra y k el número de variables independientes.

Existen varios procedimientos para la selección de las variables en dónde para el acercamiento más cercano corresponde a una búsqueda exhaustiva en el espacio de modelos. Para ello ajustamos un modelo de mínimos cuadrados para todas las combinaciones posibles de variables y seleccionamos entre ellos el modelo que mejor equilibre el tamaño y error. Sin embargo no podemos, explorar todos los modelos para p medianos y grandes dado que existen 2^p modelos posibles para p variables. (Abad, 2017-2018). Ejemplo para $p = 50$ existen más de un millón de modelos. Existen varios métodos de exploración para el espacio de modelos, en donde se mencionaran los más utilizado para el modelamiento de riesgo de crédito:

1. **"Eliminación hacia atrás" ("Backward Stepwise Regression"):** Empieza por incluir en el modelo a todas las variables regresoras disponibles y se van excluyendo de una en una según su capacidad explicativa. Es decir, la primera variable que se elimina será aquella que represente un mayor nivel de significancia que esté por encima del seleccionado, y así sucesivamente hasta que todas las variables seleccionadas se encuentren dentro del nivel de significancia en base a la prueba t o F (Fisher).

2. **"Introducción hacia adelante" ("Fordward Stepwise Regression"):** Esta metodología empieza por un modelo que no contiene ninguna variable explicativa y añade como primera de ellas a la que represente un menor

nivel de significancia que esté por encima del seleccionado. En los pasos sucesivos se va incorporando el modelo que aporte con aquella variable que represente un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento culmina cuando no es posible ingresar una nueva variable dentro del nivel de significancia seleccionado

La función step de Rstudio dentro de los modelos glm (Modelos lineales generalizados), considera ambos entradas de selección en donde para este estudio seleccionaremos la metodología backward junto con el criterio de información de Akaike (AIC, siglas en inglés). (Akaike, 1974), para asegurar la calidad del modelo dado que AIC toma en consideración el modelo que se ajuste a las series observadas y al número de parámetros utilizados en el ajuste. A menor AIC, mayor es la robustez del modelo.

2.2.3 CRITERIO DE INFORMACIÓN AKAIKE (AIC)

Cuando se tienen varias variables explicativas potenciales, las estrategias de regresión anteriores definen normalmente un subconjunto posible de modelos y el problema radica en seleccionar el mejor entre ellos.
Akaike, es la medida de ajuste de un modelo estadístico y describe la relación entre el sesgo y la varianza dentro de la construcción de un modelo, la cual viene dada por:

$$AIC = 2K - \ln(L) \quad (\mathbf{1.9})$$

En dónde, K, es el número de parámetros y L el valor máximo de la función de verosimilitud (Maximum likelihood). Por lo tanto el criterio de información Akaike evalúa la bondad de ajuste a partir de la FMV, y el número de parámetros.

La verosimilitud, $p(y|\theta_k)$ es la probabilidad condicionada de los datos a los k parámetros del modelo, siendo esta interpretada como una medida inversa a la sumatoria de las distancias euclídeas para cada uno de los datos al modelo, por ende, existe un único conjunto de valores de los parámetros que maximizan la función de verosimilitud minimizando la sumatoria de las distancias euclídeas de

los datos al modelo. Por lo tanto, cuanto mayor sea el valor de la máxima verosimilitud mejor será el ajuste del modelo.

La medida de bondad de ajuste en este caso es $2\ln(L)$. Esta es una función creciente debido a que la bondad de ajuste aumenta de forma logarítmica respecto al valor de máxima verosimilitud, L, pero se la denota como decreciente al multiplicarle por menos 1. Así, para que cuando L tienda al infinito, la expresión $-2\ln(L)$ que aparece en la función del criterio de información Akaike sea a un valor muy pequeño (el indicador AIC disminuye cuando incrementa la bondad de ajuste). Referente a la máxima verosimilitud, L, resulta del producto de las probabilidades de cada dato condicionado al modelo, por lo que L se obtiene multiplicando n valores entre cero y uno (o sumatoria de los logaritmos). Por lo tanto, L depende tanto de la distancia de cada dato al modelo como del número de datos n.

La complejidad en el AIC viene dada por K, que representa el número de parámetros del modelo. La penalización de $2K$ del criterio de información Akaike, es similar a realizar la validación cruzada del modelo dejando un dato fuera (leave one out-cross validation). (Rubalcaba, s.f.)

Para el caso de incrementos de parámetros, la verosimilitud del modelo no cambia puesto que el AIC aumentara al orden $2\Delta K$ y si en forma viceversa lo quitamos parámetros para simplificar el modelo, podemos establecer como criterio dejar de quitar variables cuando el decremento de AIC < 2, lo que indica cambios relativamente no significativos en la bondad de ajuste para el modelo.

El criterio AIC tiene una forma menos restringida que otros indicadores de medir la complejidad del modelo: sumarizando al indicador un número entero k que corresponde al número de parámetros. El problema de utilizar k como única medida de complejidad es que su efecto es mínimo comparado con el valor que toma $-2\ln(L)$ que depende del número de datos n. Por ende no importan que tan complejo sea modelo siempre y cuando posea muchos datos para avalar cada parámetro.

Es importante recordar que el modelo ideal no existe, pero siempre es preferible contar con modelos con la menor cantidad de variables posible, puesto que además de ser más sencillos, son más estables y por ende poseen menor sesgo y variabilidad.

2.2.4 TRANSFORMACIÓN DE LOS VALORES LOGIT

Finalmente, una vez obtenida la ecuación del modelo en base a las metodologías de selección y criterios de información establecidos, procedemos a transformar los valores logit para la obtención de los puntajes de calificación de crédito.
Un modelo de puntuación, ayudará a identificar de forma automática si un cliente es de alto riesgo o no para la entidad en base a una calificación recibida. Por ejemplo si un cliente obtiene una puntuación de 999/1000 indica que es un potencial cliente con altas probabilidades de pago y viceversa.

$$score = RND\left(\frac{1000}{1 + EXP(z)}\right)$$

2.2.4 MEDIDAS DE CALIDAD DEL MODELO

En riesgo de crédito, para medir la calidad del modelo se utilizan los siguientes estadísticos:

2.2.4.1 CURVA ROC

La curva ROC o característica operativa del receptor, es una herramienta que se utiliza para visualizar, y seleccionar clasificadores según su comportamiento. Uno de los pioneros en utilizar la técnica en aprendizaje automático o Machine Learning fue (Spackman, 1989), quien manifestó el poder de las curvas para contrastar y evaluar algoritmos dentro de sus investigaciones.

Este tipo de gráfico, será útil para evaluar la capacidad de discriminación del modelo con el objetivo de asignar correctamente a los sujetos a grupos distintos y a su vez permite identificar el punto de corte óptimo para minimizar la mala clasificación.

CONCEPTOS PREVIOS

Sea la variable aleatoria Y , que sigue una distribución Bernoulli de parámetro p que llamaremos prevalencia del evento sobre la población. En donde la variable toma los valores:

$$Y = \begin{matrix} 0; Cuando\ el\ \ sujeto\ no\ presenta\ el\ modelo \\ 1; Cuando\ el\ sujeto\ presenta\ el\ modelo \end{matrix}$$

Por tanto,

$$prevalencia = P(Y = 1); probabilidad\ de\ presentar\ el\ evento$$

Y, puesto que ambos estados forman un estado de sucesos:

$$1 - P = P(Y = 0); probabilidad\ de\ no\ presentar\ el\ evento$$

En dónde considera los siguientes puntos de partida:

- Una muestra aleatoria simple de un grupo de control que no presente el evento de interés, a los que llamaremos pagadores y una muestra aleatoria simple que lo presente, a los que llamaremos morosos, directamente una muestra aleatoria simple de la población.
- Una variable aleatoria x, que mida ciertas características en cada sujeto, cuya respuesta puede ser discreta o continua.
- Finalmente, una variable aleatoria Bernoulli $y\ (''Prueba'')$ de la que queremos estudiar su poder discriminante, la cual tomará los siguientes resultados: positivo o negativo, en función del valor de x, respecto al que denominaremos **umbral** $\boldsymbol{c}$. En donde los posibles valores a tomar son:

$$Y = \begin{matrix} Positivo = y = 1\,, & si\ \geq c \\ Negativo = y = 0\,, & si\ < c \end{matrix}$$

Si extraemos una muestra de la población un estimador se denotaría como la siguiente razón:

$$p = \frac{nimero\ de\ morosos\ de\ la\ muestra}{cantidad\ de\ sujetos\ de\ la\ muestra}$$

Pero si componemos nuestra muestra con una extracción de un grupo de pagadores y otra extracción de un grupo de morosos dicha razón no lo será. Por

tanto, deberemos, en este caso, extraer tantos morosos, en función de pagadores, como indique la prevalencia.

Aplicando la prueba, a morosos y pagadores se procede a dividir la población en cuatro subgrupos como lo presenta la siguiente tabla de contingencia:

Tabla 1: Matriz de confusión

	Moroso ($Y = 1$)	Moroso ($Y = 0$)
Prueba + $\equiv (y = 1)$	Verdadero positivo ($V+$)	Falso Positivo ($F+$)
Prueba - $\equiv (y = 0)$	Falso Negativo ($F-$)	Verdadero negativo ($V-$)

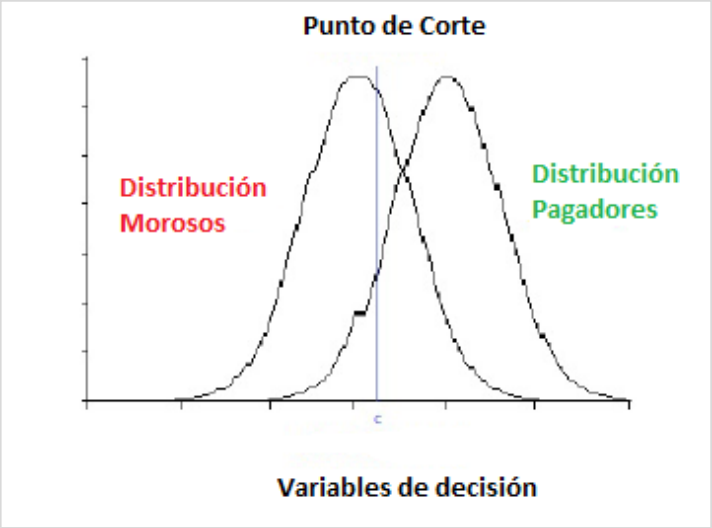
Fuente: Elaboración propia

En dónde, de manera equivalente un sujeto será:

$$V+; si(Y = 1, y = 1)$$
$$V-; si(Y = 0, y = 0)$$
$$F+; si(Y = 0, y = 1)$$
$$F-; si(Y = 1, y = 0)$$

Supongamos, que el resultado de la prueba es una variable aleatoria continua, no tiene porqué poseer la misma distribución tanto en los grupos de morosos y no morosos. De hecho, en el caso límite de que la distribución y los parámetros fuesen iguales, significaría que el comportamiento sea el mismo tanto en sujetos morosos y no morosos lo que implica que no exista el poder de discriminación dentro del modelo.

Figura 5: Distribución de Gaussiana de Morosos vs Pagadores



Fuente: Elaboración Propia

En la figura 5, podemos evidenciar en el eje de las ordenadas, la distribución Gaussiana solapada que podrían tener ambos grupos de estudio. El valor denotado como c es el punto de corte o umbral por encima del cual la prueba considerará si debe ser catalogado como sujeto moroso aunque su estado real sea pagador o viceversa.

Ante esta gráfica y la matriz de confusión lo que se espera es que ambas distribuciones estén lo más alejadas posibles, y que se minimice la presencia de falsos positivos y negativos para asegurar la potencialidad de discriminación del modelo. Para ello contamos con 2 conceptos fundamentales: Sensibilidad y Especificidad.

Es una gráfica que muestra la sensibilidad versus la especificidad en donde:

- **Sensibilidad:** Representa a la probabilidad de que a un cliente bueno la prueba de resultado sea positivo.
- **Especificidad:** Es la probabilidad de que a un cliente malo la prueba de resultado sea negativo.

Si el valor es de 0.5, entonces significaría que no existe separación. Si es 1 implica separación perfecta. Para modelos aprobación, otorgamiento y comportamiento el valor debe oscilar entre el 0.65 y 0.80.
Finalmente, la curva ROC, representa 1- especificidad frente a la sensibilidad para cada posible umbral o punto de corte.

2.2.4.2 RAZON DE ODSS

Es la relación entre clientes buenos y malos y se calcula como la razón entre el número de clientes buenos y malos. En los intervalos de score, bajo la relación de buenos decrece; y asciende exponencialmente a medida que aumenta el score.

CONCEPTOS PREVIOS

Según (Albert, 1995), la Razón Odds, es la probabilidad de que se genere un suceso dividido para la probabilidad de que no acontezca un suceso,

$$\frac{P(Y=1)}{1-P(Y=1)}$$

$$Ods_1 = \frac{p_1}{1-p_1} \qquad y \qquad Ods_2 = \frac{p_2}{1-p_2}$$

La razón de Odds se define como:

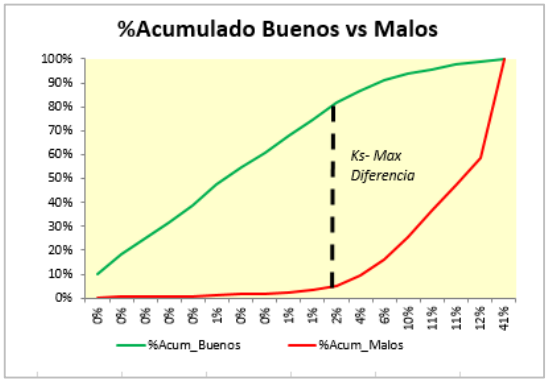
$$\boldsymbol{\theta} = \frac{Ods_1}{Ods_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

Se puede probar que para el modelo de Regresión binaria la razón Odds es la exponencial (e^{β_q}).

2.2.4.3 KOLMOGOROV-SMIRNOV (K-S):

Calcula la máxima diferencia entre dos distribuciones acumuladas, en este caso mide el poder de discriminación de los scores entre clientes buenos y malos. Entre más alto sea el coeficiente mejor será el modelo. Una tasa aceptable en modelos de aprobación u otorgamiento de crédito se aproxima al 20%, mientras que para modelos de comportamiento 40%.

Figura 6: Test Kolmogorov-Smirnov



Fuente: Elaboración Propia

2.2.4.4 ESTADÍSTICO DE GINI:

Permite evidenciar la discriminación entre buenos y malos por intervalos de score, es decir contrasta el porcentaje de clientes buenos versus los malos en los mismos intervalos o tramos de score.

Si el valor del estadístico de GINI, es de 0% significa que no distingue la separación entre buenos y malos. Un porcentaje aceptable en modelos de aprobación u otorgamiento de crédito se aproxima al 25%, mientras que para modelos de comportamiento 80%, más sin embargo se describe por intervalos los valores expresados para el coeficiente de GINI:

Figura 7: Indicadores GINI

Gini	Calidad de Clasificación
Por debajo de 25%	Baja
Entre el 25%-44%	Promedio
Entre el 45%-60%	Buena
Mas del 60%	Alta

Fuente: Elaboración Propia

2.2.5 DEFINICIÓN DE PUNTO DE CORTE

El punto de corte ayudará a definir, la línea de separación entre buenos y malos. El modelamiento en Rstudio, a su vez permite obtener el área bajo la curva de ROC, bajo la definición del **umbral** c. En donde los posibles valores a tomar son:

$$Y = \begin{matrix} Positivo = y = 1\,, & si\ \geq c \\ Negativo = y = 0\,, & si\ < c \end{matrix}$$

En este caso, utilizaremos el valor óptimo c mediante la ejecución de un algoritmo adicional que minimice el error de clasificación, dado que Rstudio considera por default el umbral de separación 0.5.

Adicionalmente, se definirán los puntos de corte para los indicadores de calidad restantes Kolmogorov Smirnov y Gini en base a los valores logit obtenidos mediante la siguiente transformación:

$$z = Logit\ (P) = ln\left(\frac{p}{1-p}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_1 + \boldsymbol{\beta_2} x_2 + \cdots + \boldsymbol{\beta_q} x_q$$

$$score = RND\left(\frac{1000}{1 + EXP(z)}\right)$$

A los valores logit transformados mediante la connotación matemática expuesta se los conoce como score, en donde en base a los puntajes obtenidos son agrupados ventiles o en veinte partes iguales para obtener la respectiva distribución homogénea por cada tramo. (Ver figura 8)

Finalmente para la obtención del punto de corte Kolmogorov Smirnov y Gini se considerarán los siguientes pasos a seguir de acuerdo a (Iñiguez & Bambino):

Pasos:

1. Ordenar los tramos de score, $\boldsymbol{k}$, de manera ascendente.
2. Determinar la proporción de buenos y malos que comparten el mismo punto de corte, $\boldsymbol{P_b(k)}$ y $\boldsymbol{P_m(k)}$, siendo:

$$\boldsymbol{P_b(k)} = \frac{\boldsymbol{n_{11}}^{\boldsymbol{k}} + \boldsymbol{n_{12}}^{\boldsymbol{k}}}{\sum_{\boldsymbol{k}} \boldsymbol{n_{11}}^{\boldsymbol{k}} + \boldsymbol{n_{12}}^{\boldsymbol{k}}} \quad \boldsymbol{y}\ \boldsymbol{P_m(k)} = \frac{\boldsymbol{n_{11}}^{\boldsymbol{k}} + \boldsymbol{n_{12}}^{\boldsymbol{k}}}{\sum_{\boldsymbol{k}} \boldsymbol{n_{11}}^{\boldsymbol{k}} + \boldsymbol{n_{12}}^{\boldsymbol{k}}}$$

1. Determinar la tasas acumulada de buenos y malos $\boldsymbol{P_b(k)}$ y $\boldsymbol{P_m(k)}$:

$$\boldsymbol{P_b(k) = \sum_{K<k} P_b(k) \;\; y \;\; P_m(k) = \sum_{K<k} P_m(k)}$$

2. Determinar las diferencias entre tasas acumuladas por puntos de corte entre buenos y malos:

$$\boldsymbol{|P_m(k) - P_b(k)|}$$

3. Identificar el punto de corte $\boldsymbol{k^*}$ que proporcione la máxima diferencia absoluta del coeficiente K-S:

$$\boldsymbol{K - S: \max_k \{P_m(k) - P_b(k)\}}$$

4. Mientras que en índice de GINI, se puede calcular a partir de la siguiente expresión:

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{n} \left(P_m(k_i) - P_m(k_{i-1})\right) * \left(P_b(k_i) - P_b(k_{i-1})\right),}$$

$$\boldsymbol{P_m(k_0) = 0, P_b(k_0) = 0,}$$

Donde:

$\boldsymbol{P_m(k_i)}$: Porcentaje acumulado de malos para un score $\boldsymbol{k_i}$

$\boldsymbol{P_m(k_{i-1})}$: Porcentaje acumulado de malos para el score anterior a $\boldsymbol{k_i}$

$\boldsymbol{P_b(k_i)}$: Porcentaje acumulado de buenos para un score $\boldsymbol{k_i}$

$\boldsymbol{P_b(k_{i-1})}$: Porcentaje acumulado de buenos para el score anterior a $\boldsymbol{k_i}$

En el ejemplo del ejercicio de la figura 8, tenemos que el punto de corte en el cual un cliente entrará en incumpliendo de pago, se encuentra a partir del tramo [860-869], es decir, si su puntaje es inferior a 860, son altas probabilidades de que el cliente no cumpla con sus obligaciones de pago.

Si existe discriminación y ordenamiento en el modelamiento, el score ayudará a identificar si los clientes que poseen mayor puntaje son buenos pagadores y viceversa. En este caso un cliente es buen pagador si su score o puntaje es mayor e igual a 870/1000. En el capítulo 3 de análisis de resultados obtendremos las tablas performance para el modelo I y II en base a la técnica propuesta.

Figura 8: Tramos de Score vs tasas de buenos y malos

Resumen por Tramos de Score

	DEFINIC								
	No.		%Columna		%Fila		Total No.tal	%Colum	Total %Fila
SCORE	BUENOS	MALOS	BUENOS	MALOS	BUENOS	MALOS			
[>=958]	146	28	7%	3%	84%	16%	174	5%	100%
[944-957]	126	27	6%	3%	82%	18%	153	5%	100%
[934-943]	130	39	6%	4%	77%	23%	169	5%	100%
[926-933]	127	33	6%	3%	79%	21%	160	5%	100%
[918-925]	118	45	5%	4%	72%	28%	163	5%	100%
[910-917]	109	48	5%	5%	69%	31%	157	5%	100%
[903-909]	106	45	5%	4%	70%	30%	151	5%	100%
[895-902]	114	49	5%	5%	70%	30%	163	5%	100%
[886-894]	120	50	6%	5%	71%	29%	170	5%	100%
[878-885]	116	44	5%	4%	73%	28%	160	5%	100%
[870-877]	116	40	5%	4%	74%	26%	156	5%	100%
[860-869]	102	55	5%	5%	65%	35%	157	5%	100%
[849-859]	101	50	5%	5%	67%	33%	151	5%	100%
[834-848]	106	66	5%	6%	62%	38%	172	5%	100%
[819-833]	105	45	5%	4%	70%	30%	150	5%	100%
[802-818]	87	69	4%	7%	56%	44%	156	5%	100%
[774-801]	100	62	5%	6%	62%	38%	162	5%	100%
[747-773]	78	85	4%	8%	48%	52%	163	5%	100%
[<=746]	147	168	7%	16%	47%	53%	315	10%	100%
Total general	**2.154**	**1.048**	**100%**	**100%**	**67%**	**33%**	**3.202**	**100%**	**100%**

Fuente: Elaboración Propia

Figura 9: Puntos de corte K-S y GINI

Resumen por Tramos de Score

Punto Corte	Buenos	Malos	Total general	%Acum_Buenos	%Acum_ Malos	K-S	GINI		
[>=958]	7%	3%	5%	7%	3%	4%	7%	3%	0%
[944-957]	6%	3%	5%	13%	5%	7%	6%	8%	0%
[934-943]	6%	4%	5%	19%	9%	10%	6%	14%	1%
[926-933]	6%	3%	5%	25%	12%	12%	6%	21%	1%
[918-925]	5%	4%	5%	30%	16%	14%	5%	29%	2%
[910-917]	5%	5%	5%	35%	21%	14%	5%	37%	2%
[903-909]	5%	4%	5%	40%	25%	15%	5%	46%	2%
[895-902]	5%	5%	5%	45%	30%	15%	5%	55%	3%
[886-894]	6%	5%	5%	51%	35%	16%	6%	65%	4%
[878-885]	**5%**	**4%**	**5%**	**56%**	**39%**	**17%**	**5%**	**74%**	**4%**
[870-877]	**5%**	**4%**	**5%**	**62%**	**43%**	**19%**	**5%**	**82%**	**4%**
[860-869]	**5%**	**5%**	**5%**	**66%**	**48%**	**18%**	**5%**	**91%**	**4%**
[849-859]	5%	5%	5%	71%	53%	18%	5%	101%	5%
[834-848]	5%	6%	5%	76%	59%	17%	5%	112%	6%
[819-833]	5%	4%	5%	81%	63%	18%	5%	122%	6%
[802-818]	4%	7%	5%	85%	70%	15%	4%	133%	5%
[774-801]	5%	6%	5%	90%	76%	14%	5%	146%	7%
[747-773]	4%	8%	5%	93%	84%	9%	4%	160%	6%
[<=746]	7%	16%	10%	100%	100%	0%	7%	184%	13%
Total general	**100%**	**100%**	**100%**			**19%**			**74%**

Fuente: Elaboración Propia

2.3 ÁRBOLES DE CLASIFICACIÓN

Los árboles de decisión, constituyen unas de las técnicas más populares en data mining por su contribución en análisis clasificación y regresión. De entre las diversas metodologías existentes para construir árboles de clasificación, utilizaremos la metodología de árboles de clasificación y regresión (CART), propuesta por (Breiman, Friedman, Olshen, & Stone, 1984) e implementada dentro en la libraría de rpart de Rstudio por los autores: (Therneau, Atkinson, & Bryan,Rypley, 2017).

CONCEPTOS PREVIOS

Sea Y una variable de respuesta cualitativa con K clases, modalidades o grupos $C_1, \dots, C_K$ y sea $X = X_1, \dots, X_p$, p variables predictoras. Lo que se busca es establecer una relación entre Y variables de respuesta y las $X_1, \dots, X_p$ variables predictoras, es decir, determinar una función objetivo f tal que $Y \approx X_1, \dots, X_p$, de modo que dado un sujeto o individuo de estudio se pueda asignar a una de las clases con el menor error posible, para así identificar la regla de decisión final con el mayor de aciertos posible.

El crecimiento de un árbol de clasificación se realiza mediante la técnica de división binaria recursiva, de esta forma se tendrá un conjunto de M nodos terminales, asociados a M regiones $\{R_1, \dots, R_M\}$ que conforman una partición del espacio predictor, para ello, se le asignará a cada nodo o región una de las clases con el objetivo de realizar una predicción a partir de una observación dada.

A continuación se presentan las siguientes definiciones a priori para la construcción del árbol de decisión (Pino, 2017):

- n_t : Número de observaciones en el nodo t
- $n(k)$: Número de observaciones que pertenecen a la clase $\mathrm{C_k}$, $k = 1,2, \dots K$
- $\pi_k : k = 1, .. K$: Son las probabilidades a priori, y se definen como la probabilidad de que una observación pertenezca a la clase k. Si son desconocidas, de ser el caso, pueden ser estimadas mediante las frecuencias relativas: $\pi_k : \frac{n(k)}{n}$.
- $\mathrm{n_t}\ (m)$: Número de observaciones de la clase $\mathrm{C_k}$ pertenecientes al nodo t.

- $v(X)$: Clase a la que pertenece un caso cuyo vector de variables predictoras es X.
- $d(t)$: Decisión en el nodo t, para todo caso en el que pertenezca al nodo t

2.3.1 CONCEPTOS ASOCIADOS AL ÁRBOL DE DECISIÓN

A continuación se definen algunos conceptos asociados con el árbol de decisión que serán de gran ayuda para la elaboración del presente trabajo (Pino, 2017):

La probabilidad asociada al nodo t: $p[t] = P_r[X \in t]$, es la probabilidad de que una observación se encuentre en el nodo t.

Si $\{\pi_k\}$ son conocidas, se puede estimar mediante la expresión:

$$P(t) = \sum_{k=1}^{K} \pi_k \frac{n_t(k)}{n(k)}$$

Si $\{\pi_k\}$ son desconocidas, se utiliza como estimador:

$$P(t) = \sum_{k=1}^{k} \frac{n(k)}{n} \frac{n_t(k)}{n(k)} = \sum_{k=1}^{k} \frac{n_t(k)}{n} = \frac{n_t}{n}$$

En este caso, se puede evidenciar que $P(t)$ es la proporción de casos del conjunto de entrenamiento que pertenecen al nodo t.

- Probabilidad de la clase C_k dado el nodo t: $P(k|\mathrm{t}) = P_r[v(X) = k|\mathrm{X} \in \mathrm{t}]$, es decir, la probabilidad de que un sujeto pertenezca a la clase C_k sabiendo que se encuentra en el nodo t.

Si $\{\pi_k\}$ son conocidas, se puede estimar mediante la siguiente expresión:

$$P(k|t) = \frac{\pi_k \frac{n_{t(k)}}{n(k)}}{\sum_{k=1}^{k} \pi_k \frac{n_t(k)}{n(k)}}$$

Para el caso de que $\{\pi_k\}$ sean desconocidas, se utiliza como estimador:

$$P(k|\mathrm{t}) = \frac{\sum_{k=1}^{k} \frac{n(k)}{n} \frac{n_t(k)}{n(k)}}{\frac{n_t}{n}} = \frac{n_t(k)}{n_t}$$

Obsérvese que, para el último caso, $P(k|\mathrm{t})$, resulta ser la proporción de los casos de entrenamiento del nodo t que pertenece a la clase C_k. En general, el criterio de decisión para cada nodo viene dada por la clase más probable en ese nodo. Es decir, cualquier caso perteneciente al nodo t es clasificado en la clase donde alcanza el máximo de las probabilidades de las diferentes clases:

$$d(t) = argmax_{k=1\ldots K} P(k|\mathrm{t})$$

En caso de que las probabilidades sean iguales, se elige aleatoriamente una de las clases de máxima probabilidad.

El error de clasificación en el nodo t se puede evaluar a través de la probabilidad de clasificación incorrecta para el nodo t.

$$r(t) = 1 - P(d(t)|\mathrm{t})$$

El estadístico $r(t)$, se lo conoce como el estimador por re-sustitución, y en general representa la proporción de sujetos del nodo t que no pertenecen a la clase $d(t)$.

Se define el riesgo del nodo t como:

$$R(t) = p(t)r(t)$$

Con esta definición se puede ampliar al árbol completo, de modo que el riesgo del árbol T se define finalmente como el estimador por re-sustitución de la tasa de error esperada del árbol de clasificación. Suponiendo M nodos finales:

$$R(T) = \sum_{t=1}^{M} p(t)\, r(t) = \sum_{t=1}^{M} R(t)$$

2.3.2 ELEMENTOS PARA EL ALGORITMO DE CONSTRUCCIÓN

Dados los conceptos básicos definidos en el punto anterior, en esta sección se definirán los criterios necesarios para la construcción del árbol de decisión:

a) Un conjunto Ω de preguntas binarias para las segmentaciones
b) Evaluar la bondad de las divisiones
c) Una regla de parada
d) Reglas para asignar una clase a un nodo terminal

a) Un conjunto Ω de preguntas binarias para las segmentaciones

En el nodo inicial o raíz se encuentran las n observaciones de la muestra de entrenamiento. El método consiste en dividir el conjunto de aprendizaje en dos partes. Para ello se requiere de un conjunto de preguntas binarias Ω y escoger la más óptima. De esta forma se tienen dos nodos sucesores, uno a la izquierda t_L y otro a la derecha t_R .Cada una de las segmentaciones va a depender de la única variable, que en función de su categoría, tendrán diferente divisiones, de modo que:

- Si las variables binarias, son del tipo 1/0, Sí/No, etc. Sólo existiría una posible división o ;
- Si las variables son cuantitativas u ordinales, existirían infinitas divisiones del tipo:

$$\{x \leq c\} c \ \in R$$

Aunque inicialmente existan $\propto$ divisiones de este tipo, se puede considerar una cantidad finita o limitada seleccionando los **puntos medios** entre dos observaciones ordenadas de la muestra como posibles puntos de división.

- Para variables nominales con B modalidades, las posibles divisiones son:

$$\frac{[VR_2{}^B - 2]}{2} = \frac{[2^B - 2]}{2} = 2^{B-1} - 1$$

Para seleccionar la división óptima en cada nodo la idea es identificar la mejor división. Para cada variable se compararían las p mejores divisiones de las variables individuales y se selecciona la mejor entre todas ellas. Para ello, será necesario realizarlo en base algún criterio que mida la bondad de las divisiones.

b) Evaluar la bondad de las divisiones

Ubicados en el nodo t, una elección para determinar la mejor división sería seleccionar aquella que represente una mayor reducción en el riesgo $R(T)$. Este criterio puede producir árboles de bajo rendimiento, por tener numerosos nodos finales. En donde para este caso introducimos el término de funciones de impureza:

Una función de impureza Φ se define como el conjunto de todas las K-tuplas de números $(p_{1,..,}p_k)$ satisfaciendo $p_k \geq 0, k = 1 \dots K, \sum_{k=1}^{K} p_k = 1$, tomando en consideración las siguientes propiedades:

- Φ es un máximo sólo en el punto $(\frac{1}{K},..,\frac{1}{K})$
- Φ logra su mínimo solo en los puntos (1,0,..,0), (0,1,..,0),…, (0,0,..,1).
- Φ es una función simétrica de $p_{1,..,}p_k$

Dada una función de impureza Φ, se define a la medida de impureza $I(t)$ de cualquier nodo t como:

$$I(t) = \Phi(\mathrm{P}(1|\mathrm{t}), \mathrm{P}(2|\mathrm{t}), .. \mathrm{P}(\mathrm{K}|\mathrm{t}))$$

Situados, en un nodo t, al realizar la división, una parte de los datos se asignan a la rama derecha t_R con probabilidad $P(t_R)$ y otra a la rama izquierda t_L con probabilidad $P(t_L)$, de forma que, la disminución de la impureza en el nodo t es:

$$\Delta\mathrm{I}(\mathrm{t}) = \mathrm{P}(\mathrm{t})\mathrm{I}(\mathrm{t}) - P(t_R)I(t_R) - P(t_L)I(t_L) \geq 0$$

Por tanto, para determinar la división del nodo t, se elegirá la división en la que la variable maximice la reducción de la impureza $\Delta\mathrm{I}$.

Existen numerosas funciones de impureza, pero las más habituales son el índice de diversidad de Gini y la entropía. (Hastie, Tibshirani, & Friedman, 2009)

- **Índice de diversidad de Gini:** Dado $(p_{1,..,}p_k)$ se define el índice de diversidad de Gini como:

$$\Phi(p_{1,..,}p_k) = \sum_{k \neq j} p_k p_j = 1 - \sum_{k=1}^{K} {p_k}^2$$

- **Entropía:** Dado $(p_{1,..,}p_k)$ se define su entropía como:

$$\Phi(p_{1,..,}p_k) = -\sum_{k \neq j} p_k \log(p_k)$$

  En el caso en que $p_k = 1$, se considera 0, $\log(1) = 0$

Una vez realizadas todas las consideraciones para cada criterio, se procede a la construcción del árbol. Para ello, se deben seguir los siguientes pasos:

- Obtener el nodo raíz, que incluya todos los individuos del conjunto de entrenamiento.
- Obtener el par (s, t) donde s es (variable, división), es decir, la variable y el punto por donde se realiza la división y t es el nodo donde la queremos efectuar a cabo.
- Aplicar a cada nodo el punto anterior, hasta que se verifiquen las condiciones de finalización, en base a las reglas de parada.

c) Una regla de parada

Esta regla indicará las condiciones que se tienen que efectuar para que un nodo se divida y por tanto sea terminal. Inicialmente se podría obtener un árbol con un nodo final por cada observación, pero se tendría un modelo resultante muy complejo, que aunque se tiene $R(t) = 0$, la capacidad de generalización del modelo sería baja y esto implicaría presencia de sobreajuste.
Para evitar el sobreajuste se deben considerar los siguientes criterios de parada:

1. Mínimo número de casos que debe tener un nodo para intentar segmentar en dos nodos hijos.
2. Mínimo número de casos en un nodo terminal

3. Las particiones deben producir una reducción mínima de la función de impureza. Es decir, establecemos un umbral $\beta > 0$, para lo cual se declara un nodo terminal t si:

$$max_{s \in S} \Delta I(s,t) < \beta$$

Donde S es el conjunto de todos los pares (variable, división) que se han elaborado en el árbol.

Supóngase, que se han realizado algunas particiones y hemos llegado a un conjunto de nodos terminales. El conjunto de divisiones usadas, junto al orden en que hemos hecho las divisiones determina lo que llamamos un árbol binario T.

d) Reglas asignar una clase a un nodo terminal

Finalmente, para la regla de asignación, se seleccionará una clase que verifique:

Si $x \in R_t$ se asigna a la clase $C_d(t)$ para la que se alcanza: $max_{k=1..K} p(\mathrm{k}|\mathrm{t})$

El crecimiento excesivo del árbol, puede dar paso a que la data inicial de entrenamiento se ajuste bien al modelo, pero esto no implica que se puedan producir los mismos resultados para los diferentes datos de entrada, por ende para evitar fallas en el método se introduce el concepto de poda o pruning.

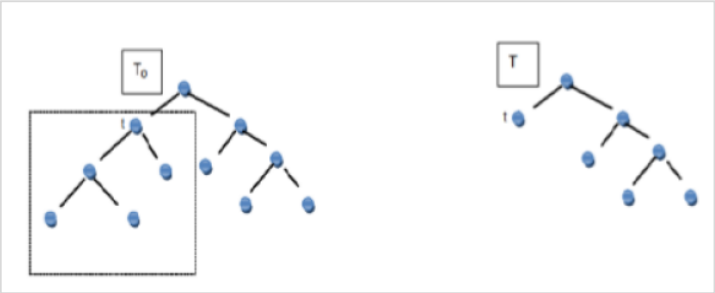
2.3.3 PODA DEL ÁRBOL (PRUNING)

El número de nodos terminales o el tamaño del árbol, es un parámetro que controla la complejidad del modelo, dado que si un árbol es muy pequeño por ejemplo, puede no capturar la estructura de los datos, mientras que si es muy grande puede sobreajustar los datos.

Tomando, en consideración el apartado anterior, una vez construido el árbol T_0, se procede con el pruning del mismo, seccionando sucesivamente ramas o nodos terminales que constituyan poca contribución a la explicación de la variable de respuesta, para así encontrar el tamaño óptimo del árbol.

Como paso inicial, hay que construir un árbol de tamaño muy grande T_0 permitiendo que el criterio de seccionamiento continúe hasta que la regla de parada del criterio de crecimiento del árbol determine finalmente los nodos terminales. Finalmente, se procede al pruning que parte de un nodo t. El pruning del árbol consiste en eliminar de T_0 todos los sucesores de dicho nodo, esto es, excluir todo lo que esté por debajo del nodo t, tal como se evidencia en la figura 10 ; para este caso el árbol inicial T_0 es el de la izquierda y el árbol podado T el de la derecha. Si se obtiene un subárbol T a partir de T_0 por sucesivas podas de ramas, entonces T es llamado subárbol podado de T_0 y denotado por $T \subseteq T_0$. Tomar en consideración que T_0 y T tienen el mismo nodo raíz.

Figura 10: Poda de Árbol



Fuente: Pino,2017

2.3.4 SECUENCIA DE SUBÁRBOLES

Dado a que un árbol se puede seccionar o podar para cada uno de sus nodos de forma anidada, existe una selección de subárboles de dimensión cada vez más pequeños.

Una vez creada la secuencia, se tiene que elegir cuál es el mejor subárbol. Para ello se define el criterio de coste-complejidad, el cual está basado en considerar una secuencia de árboles indexada para un parámetro de ajuste α no negativo. A cada α le corresponde un árbol $T_\propto \subseteq T_0$ tal que:

$$C(t_\propto, \propto) = min_{T \subseteq T_0} C(T, \propto) \; siendo \; C(T, \propto) = R(T) + (\propto |\mathrm{t})$$

Donde $|T|$ es el tamaño de T y $R(T)$ el riesgo del árbol lo cual podría sustituirse por la función impureza para la creación del árbol.

El parámetro α controla el compromiso entre el tamaño del árbol y el ajuste a los valores dados. Valores pequeños de α conducen a un tamaño del árbol grande y viceversa. Para cada valor de α, se encuentran subárboles $T_\propto \subseteq T_0$ que se minimizan $C(t_\propto, \propto)$. Cuando α= 0, entonces el subárbol $T_\propto$ será igual a T_0, puesto que el árbol coincide con el inicial.

A pesar de que los valores de α son infinitos, el número de subárboles es finito y se denotará por m. Esto es debido a que el intervalo se encuentra entre $[\propto_h, \propto_{h+1})$ en donde el árbol óptimo es el mismo. De hecho, se puede obtener una familia anidada de subárboles $\{T^{(h)}\}_h$ y una serie de valores $\propto_1 \ldots \propto_{m-1}$ tales que $T^{(h)} = T_\propto$ para todo $\propto \in (\propto_h, \propto_{h+1}),\ h = 1 \ldots m-1$.

2.3.5 SELECCIÓN DEL ÁRBOL ÓPTIMO

Finalmente, una vez obtenida la secuencia de subárboles, debe optarse por uno de ellos. A continuación, se debe disponer de un estimador insesgado del error esperado para cada subárbol.
El método definido anteriormente crea una sucesión decreciente de subárboles anidados $T_0 \supseteq T_{1\ldots} \supseteq t_1$, donde t_1 es el nodo raíz, es decir el mínimo subárbol posible de T_0.

El consiguiente árbol que se plantea consiste en la selección del árbol óptimo. Para ello, se asocia una medida de error a cada árbol y se opta por aquel que tenga asociado un mínimo error.

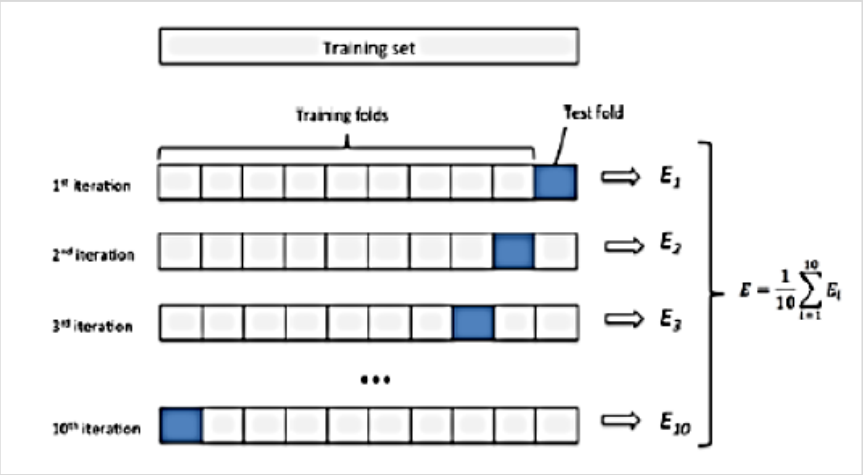
$$\text{Se elige } \ T^{(k_0)}: C\left(T^{(k_0)}\right) = minC\left(T^{(k)}\right) conC(T) = C(T, 0).$$

Por lo tanto, se debe disponer de un estimador insesgado del error (riesgo) esperado para cada subárbol. Este estimador se obtiene mediante la validación cruzada.

Para la construcción de la metodología de validación cruzada de k iteraciones (o k – fold cross validation), $C_{CV}(T)$, se divide la muestra de aprendizaje L en k subconjuntos. Generalmente se considera el número de iteraciones $k = 10$, de modo que, para la primera iteración, se entrene un modelo par $k - 1$ subconjuntos, dejando fuera el primero, para posteriormente evaluarlo, en donde este proceso se lleva a cabo k veces.

En la figura 11, se muestra lo expuesto anteriormente, en donde se realiza un proceso de validación cruzada para 10 iteraciones, y para cada conjunto el test fold va cambiando con el objetivo de realizar una estimación para cada caso para luego obtener una estimación final a través de la ponderación individual.

Figura 11: Esquema de Validación Cruzada



Fuente: Pino,2017

Ya una vez que se posee cada uno de los subconjuntos evaluados para las diferentes iteraciones, se calcula la media global de las k medidas obtenidas a través de las iteraciones, que denotaremos como $C_{CV}(T)$. Es habitual utilizar este método cuando la muestra no es lo suficientemente grande.

A pesar de los avances o mejoras introducidas en esta técnica, la desventaja de este algoritmo es que es muy inestable, es decir, debido a que una pequeña variación en el conjunto de datos conlleva a la generación de árboles distinto.

El estimador $C(T)$ en función del número de nodos terminales $\left|T^{(k)}\right|$, se comporta, de acuerdo a la figura 12, donde se pueden observar tres zonas diferenciadas: un decrecimiento inicial, a continuación una zona relativamente constante, y por último crece para valores más incrementales de $\left|T^{(k)}\right|$.

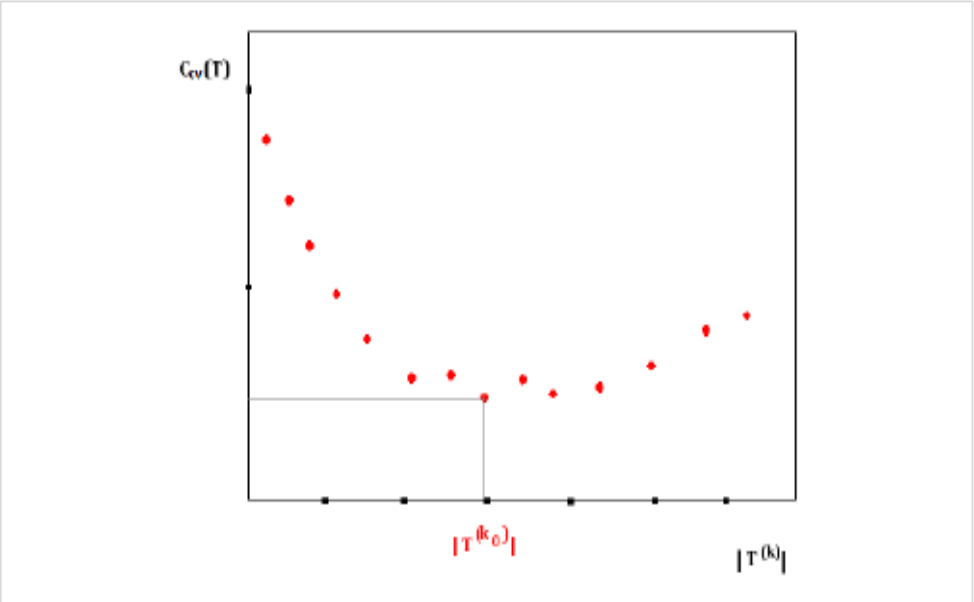
El método de validación cruzada tiene una variante o riesgo debido a que la estimación se hace en función del conjunto de entrenamiento y para evitar la inestabilidad del estimador es recomendable ajustar la incertidumbre de $C_{CV}(T)$ mediante el cálculo del error estándar (SE) . El error estándar asociado al estimador $C_{CV}(T)$, viene dado por:

$$SE(C_{CV}(T)) = \sqrt{C_{CV}(T)\frac{1 - C_{CV}(T)}{|L_T|}}$$

Donde L_T es el subconjunto de los datos de la muestra del test utilizados para construir el árbol T.

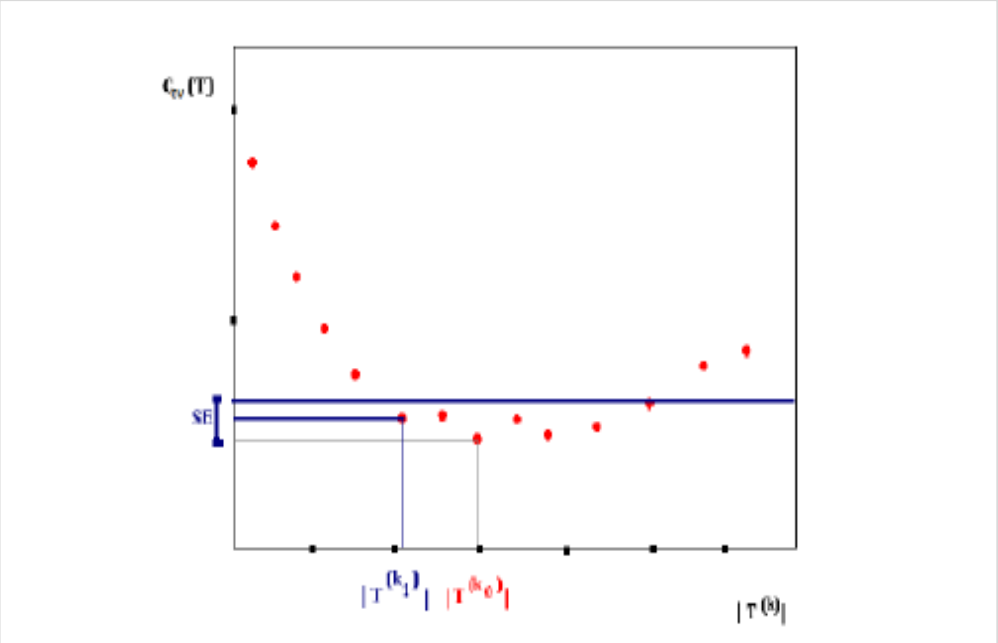
En la figura 12, se puede evidenciar que el valor mínimo es muy inestable. Una solución alterna es seleccionar $1 - SE(1 - error\ estandar)$ con el objetivo de contrarrestar esta inestabilidad y a la vez asegurar que el árbol seleccionado sea el más simple con valor cercano mínimo. En la figura 13 se observa que el mínimo $C_{CV}(T)$ se alcanza a k_0 . A partir del mismo, se calcula $SE(C_{CV}(T^{k_0}))$ y se elige aquel punto que este por debajo de la recta de SE con el menor número de nodos posibles, en este caso es k_1.

Figura 12: Estimador de $C_{CV}(T)$ en función de los nodos terminales



Fuente: Pino,2017

Figura 13: Estimador de $C_{CV}(T)$ en función de los nodos terminales con el error Estándar SE



Fuente: Pino,2017

De forma que la regla $\mathbf{1} - \boldsymbol{SE}$ se define como:

Si $\boldsymbol{T^{k_0}}$: $\boldsymbol{C_{CV}(T^{k_0}) = min_k C_{CV}(T^k)}$ entonces se selecciona $\boldsymbol{T^{(k_1)}}$ con $\boldsymbol{k_1}$ tal que:

$$\boldsymbol{cp = k_1 = max\{k: C_{CV}(T^k) \leq C_{CV}(T^{k_0}) + SE(C_{CV}(T))\}}$$

Por tanto, se elige el árbol más simple, que será el que tenga el menor número de nodos terminales y que no se desvíe más de la regla $\mathbf{1} - \boldsymbol{SE}$ del mínimo.

2.4 BAGGING

Los métodos de combinación de modelos resultan básicamente de la agregación de cierto número de modelos para obtener una clasificación o predicción a partir de los diferentes clasificadores o predictores previamente generados.

La construcción de cada modelo puede depender de los siguientes aspectos (Pino, 2017):

- Definición del conjunto de entrenamiento como por ejemplo muestras, reponderaciones, o bootstrap.
- Elección de variables.
- Selección del modelo.

Los modelos Random Forest o bosques aleatorios se basan en las técnicas bagging o ensacado. El termino *bagging* viene dado por *"Bootstrap aggregating",* la técnica fue propuesta por (Breiman L. , Machine Learning:Bagging Predictors, 1996).

Un concepto esencial en este contexto es la de la técnica de remuestreo o bootstrap, la cual resulta de la obtención de una muestra aleatoria de tamaño n extraída con reemplazamiento a partir de la muestra disponible.

A continuación se define el procedimiento *bagging* basado en problemas regresión y clasificación.

Sea un conjunto de datos $D = \{D_i = (X_i, Y_i), i = 1,2, \dots, n\}$ y un modelo de predicción $g(x)$ para $E[Y|X = x]$, con X p-dimensional y Y real. Se define al predictor *bagging* como el valor esperado $E[\hat{g}^*(x)]$, del predictor evaluado sobre muestras bootsrap (Pino, 2017) :

$$\hat{g}_{bag}(x) = E[\hat{g}^*(x)] = E[\hat{g}(x; D^*{}_1, \dots, D^*{}_n)],$$

El predictor requiere generar todas las muestras bootstrap posibles, lo cual puede ser inviable en términos computacionales, donde para ello se obtiene una

proximidad mediante la generación aleatoria de un número B de muestras bootstrap.

Algoritmo bagging (Regresión)

Para $b = 1,2 \dots, B$

1. Generar una muestra bootstrap
2. Calcular $\hat{g}(x; {D^*}_1{}^b, \dots, {D^*}_n{}^b)$
3. Aproximar el predictor bagging mediante $\hat{g}_{bag,B}(\mathrm{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{g}(x; {D^*}_1{}^b, \dots, {D^*}_n{}^b)$

En problemas de clasificación el clasificador agregado se calcula a partir de la votación de los B clasificadores generados:

$$({g^*}_1, \dots, {g^*}_k)\ {g^*}_j = \frac{\sum_{b=1}^{B} \hat{g}(x; {D^*}_1{}^b, \dots, {D^*}_n{}^b)}{B}$$

$$\hat{g}_{bag,B}(\mathrm{x}) = \mathrm{argmax}_j {g^*}_j$$

Una opción es trabajar con las estimaciones de las probabilidades y considerar para cada clase las media de las B probabilidades así calculadas para finalmente tomar la clase máxima probabilidad media. El bagging tiende a disminuir la varianza del predictor, sobre todo con modelos poco estables, como árboles de decisión o las redes neuronales artificiales (Breiman L. , Machine Learning:Bagging Predictors, 1996):

$$var = \{\hat{g}_{bag}(x)\} \approx\leq var\{g(x)\}$$

Como dificultad presenta cierta propensión a aumentar el sesgo cuadrado, pero no impide la disminución del error cuadrático medio.

La estructura de las muestras bootstrap permiten adquirir un estimador insesgado del error de predicción aunque no se disponga del conjunto test. Este estimador, es conocido como estimador OBB ("Out of Bag") o fuera del saco, la cual se basa en el aprovechamiento de las observaciones no contenidas en cada muestra bootstrap.

En un problema de clasificación con K clases se obtiene de la siguiente forma:

Sea un conjunto de entrenamiento $D = \{D_i = (X_i, Y_i), i = 1..n\}$, donde las clases de cada caso se identifican mediantes las Y_i.

Algoritmo:

Para $b = 1$ hasta B

1. Generar una muestra bootstrap D^* del conjunto D.
2. Sea $D_b = \{D_i / D_i D^*\} = D - D^*$
   2.1 Construir el modelo A_b sobre D^*
   2.2 Aplicar A_b a cada elemento de D_b
3. Siguiente b

   3.1 Para cada caso D_i se consideran las predicciones para dicho caso proporcionadas por aquellos modelos en cuyo conjunto de entrenamiento no se incluye D_i.
   3.2 De forma similar al procedimiento de validación cruzada, se obtiene la predicción agregada para D_i mediante la clase donde más veces es clasificado dicho caso por los modelos indicados en el párrafo anterior.
4. Se define el estimador OOB:

$$OBB = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq V_i)$$

Por tanto la tasa de error de Out of Bag, es la proporción de casos cuya clasificación más frecuente no coincide con su clase real, entre los modelos ajustados sobre muestras bootstrap que no lo contienen.

2.5 BOSQUES ALEATORIOS

El algoritmo de bosques aleatorios o Random Forest fue propuesto por **(Breiman L. , 2001)**, a partir de la metodología bagging. Rstudio ofrece el paquete randomForest en donde gracias a la contribución de los autores **(Liaw & Wiener, 2002)** , podemos obtener la iteración de alrededor de 500 árboles disponibles para el contexto de este análisis.

Para la construcción de cada árbol, se obtuvieron un conjunto de datos de entrenamiento de tamaño $\boldsymbol{n}$ con $\boldsymbol{p}$ variables predictoras, las cuales se desarrollan de acuerdo a los siguientes pasos **(Pino, 2017)**:

1. Seleccionar una muestra con reemplazamiento de tamaño n de la muestra de entrenamiento (bootstrap).
2. En cada nodo del árbol construido para cada muestra bootstrap, se eligen aleatoriamente $m < p$ variables predictoras, y se elige la mejor división entre esas m variables.
3. Cada árbol se construye hasta alcanzar un tamaño razonablemente grande, sin realizar poda.

En la librería randomForest de Rstudio por defecto, $m = p^{1/2}$ (problemas de clasificación) o, $m = \frac{p}{3}$ (problemas de regresión).

La tasa de error del modelo final depende de los elementos:

- La correlación entre dos árboles cualesquiera del bosque. A mayor correlación menor error.
- La fuerza de cada árbol en el bosque. Un árbol con una tasa de error reducida es un clasificador fuerte. Aumentar la fuerza de los árboles individuales disminuye la tasa de error del bosque.

Al reducir m reduce tanto la correlación como la fuerza, y viceversa. Este es el único parámetro a ajustar respecto al cual Random Forest es sensible, puede ser ajustado con la ayuda de procedimiento de validación cruzada.

Para finalizar este apartado se describe el algoritmo que sigue la técnica Random Forest para medir la importancia de cada variable.

Sean p variables predictoras:

Para $b = 1$ hasta B

1. Generar una muestra bootstrap D^* del conjunto D.
2. Sea $D_b = \{D_i / D_i D^*\} = D - D^*$
   2.1 Construir el modelo A_b sobre D^*
   2.2 Aplicar A_b a cada elemento de D_b
3. Siguiente b

2.6 CADENAS DE MARKOV

Las cadenas de Markov constituyen un apartado muy importante para el análisis y tratamiento de problemas de característica aleatoria. Andrei Markov (1856-1922), creador de las Cadenas de Markov, sostuvo que existen ciertos procesos cuyo estado futuro depende de su estado actual y que a su vez son independiente a sus estados pasados. Esta metodología resulta ser una herramienta muy importante en el contexto de modelos de riesgo de crédito, dado que permite obtener las categorías de la variable dependiente de estudio en base a los apartados de matrices de transición y roll rate que se verán a continuación.

CONCEPTOS PREVIOS

Suponiendo que $\boldsymbol{n}$, es el estado actual del proceso y sus estados anteriores son conocidos, entonces la probabilidad de todos sus estados fututos, $\boldsymbol{X_1, X_2 \ldots X_{n+1}}$ dependerán únicamente del estado actual $\boldsymbol{X_n}$, y no de los anteriores $\boldsymbol{X_1, X_2 \ldots X_{n-1}}$. **(Rojo & Miranda, 2009)**

DEFINICIÓN FORMAL

El proceso estocástico[2], $\{X_n \; con \; n = 0,1,2 \ldots\}$ si para cada i, j, n se tiene:

$$\boldsymbol{P\{X_n = j / X_{n-1} = i\} = p_{ij}}$$

Donde $\boldsymbol{p_{ij}}$, es la probabilidad de que $\boldsymbol{X}$ en el tiempo $\boldsymbol{n}$ sea igual a $\boldsymbol{j}$, condicionando a que $\boldsymbol{X}$ en los tiempos $\boldsymbol{0, 1, 2 \ldots n-1}$ fuese igual a $\boldsymbol{i}$.

CADENAS ESTACIONARIAS Y NO ESTACIONARIAS

Para determinar si una cadena de Markov, depende del tiempo o no se establecen las siguientes definiciones de cadenas estacionarias y no estacionarias:

- **Cadena Estacionaria u Homogénea:** Si las probabilidades de transición son independientes del tiempo.
- **Cadena no Estacionaria u no Homogénea:** Si las probabilidades de transición son dependientes del tiempo.

Entre ambas definiciones las más aplicada en el mundo real, son las cadenas de Markov Estacionaria, para el caso de transición de días de mora.

ESTADOS DE UNA CADENA DE MARKOV

Los estados de una cadena de Markov, se pueden definir de la siguiente forma:

- **Estado Absorbente:** Cuando la probabilidad de seguir en el mismo estado es del 100%, es decir una vez que se ingresa en el ya no se puede salir.
- **Estado Recurrente:** Cuando partiendo de un estado inicial existe la probabilidad de volver en algún momento del tiempo sobre sí mismo.

[2] "**Proceso Estocástico:** Es un modelo matemático que describe el comportamiento de un sistema dinámico, sometido a un fenómeno de naturaleza aleatorio. La presencia de un fenómeno aleatorio hace que el sistema evolucione según un parámetro, que normalmente es el tiempo t cambiando probabilísticamente de estado" (Rojo & Miranda, 2009)

- **Estado Transitorio:** Cuando partiendo de un estado inicial no se tiene la certeza de volver en algún momento del tiempo sobre sí mismo.
- **Estado Periódico:** Cuando partiendo de un estado inicial, solo es probable retornar a él en un número de tapas que sea múltiplo de un cierto número entero mayor que uno, caso contrario se denomina aperiódico.
- **Estado Ergódico:** Si, todos los estados son recurrentes aperiódicos y se comunican entre sí.

2.6.1 MATRICES DE TRANSICIÓN

Para la definición de las tasas de deterioro de la entidad se utilizará dos metodologías: Matrices de transición y Roll Rate.

La matriz de transición se define como la probabilidad de que un sujeto situado en el tramo inicial de días de mora ***i*** , migre al siguiente tramo final ***j*** dentro de un horizonte de tiempo o ventana de desempeño de 1 año.

Tabla 2: Matriz de Transición y Probabilidades

Matriz de Transición		Estado Final				
		1	**2**	**3**	**….**	j
Estado Inicial	**1**	p_{11}	p_{12}	p_{13}	….	p_{1j}
	2	p_{21}	p_{22}	p_{23}	….	p_{2j}
	3	p_{31}	p_{32}	p_{33}	….	p_{3j}
	….	….	….	….	….	….
	i	p_{i1}	p_{i2}	p_{i3}	….	p_{ij}

Fuente: Elaboración Propia

En donde la matriz definida en la tabla 2, se define de la siguiente forma:

- La columna izquierda i, representa el tramo inicial de días de mora del periodo de análisis.
- La fila superior j , corresponde al tramo final donde se encuentra el sujeto de análisis.
- La diagonal principal (bloque amarillo) de la matriz representa el estado absorbente en el cual se encuentra un sujeto de crédito. Es decir la

probabilidad de que un sujeto se encuentre en el mismo tramo inicial i y final j es del 100%.

- Las probabilidades que se encuentran por debajo de la diagonal principal representan a los sujetos que mejoraron su tramo de vencido y;
- Las probabilidades que se encuentran por encima de la diagonal principal representa a los sujetos que empeoraron su tramo de vencido.

La matriz de transición de probabilidades debe cumplir con los siguientes criterios:

- Todos los elementos de la matriz deben ser positivos, es decir $P_{ij} > 0$
- La suma de los elementos de cada fila debe ser igual a 1, es decir $\sum P_{ij} = 1$ para todo i.

Dónde P_{ij} se define como: $$P_{ij} = \frac{N_{i,j}}{N_i} \ \forall \ i, j$$

$\boldsymbol{N_{i,j}}$ Número de clientes que comenzaron al inicio del periodo en el tramo inicial $\boldsymbol{i}$ y terminaron al finalizar el periodo en el tramo $\boldsymbol{j}$.

$\boldsymbol{N_i}$ Número de clientes que comenzaron al inicio del periodo en el tramo inicial $\boldsymbol{i}$.

- Adicionalmente, el proceso de migración aleatoria puede representarse a través de la metodología de cadenas de Markov como una serie de eventos; es decir la probabilidad de que ocurra un evento dependerá del evento anterior y no sobre características intrínsecas del crédito ya sea de la entidad o de condiciones externas económicas.

- La ponderación es la misma para todos los periodos.

2.6.2 MATRICES ROLL RATE

Al igual que las matrices de transición, la metodología roll rate permite identificar la tasa de deterioro dentro de la entidad, analizando los días de vencido del mes de análisis respecto al mes siguiente, con el objetivo de identificar si las tasas de deterioro de días de vencido o mora se incrementan o normalizan, permitiendo así definir el punto de default.

Se define **default**, a falta de pago o a cualquier otro tipo de contravención en base a las condiciones de un préstamo dado, ya sea por el incumplimiento del plazo o monto pactado con la identidad.

2.6.3 CRITERIOS PARA LA DEFINICIÓN DE DEFAULT

Para identificar el punto de default a continuación se establece la definición de los criterios de cada segmento para ambas metodologías:

- **Clientes Buenos:** Se define clientes buenos si la tasa de avance de mora en ambas metodologías se encuentra inferior al 42%

- **Clientes Indeterminados:** Se define clientes indeterminados si la tasa de avance de mora en ambas metodologías se encuentra entre el 42% -54%

- **Clientes Malos:** Se define clientes malos si la tasa de avance de mora en ambas metodologías es superior al 54%.

Las tasas de avance de mora son consideradas por criterios de experiencia en modelos de riesgo de crédito.

2.7 SOFTWARE UTILIZADOS

2.8.1 LENGUAJE DE PROGRAMACION R - RSTUDIO

Es un software estadístico que fue creado en 1993, por **(Ihaka & Gentleman, 1996)** , para el tratamiento, la visualización, modelización de datos y exploración estadística. El lenguaje de programación R forma parte del proyecto GNU del premiado leguaje S, el cual fue desarrollado por AT&T Bell laboratorios y se presenta como software libre, es decir, que los usuarios poseen la libertad de copiar, distribuir, estudiar, cambiar y mejorar el rendimiento de sus algoritmos sin costo alguno. Su instalación se ejecuta a través del CRAN (Comprehensive R Archive Network).

La mayoría de las técnicas estadísticas se encuentran en la base de R, pero existen otras metodologías que se encuentran como paquetes (packages). Un paquete (package), es una recopilación de funciones, datos y código que se almacenan en forma de carpeta dentro de la estructura de R. A enero 2018 existen más de 12.000 paquetes que han contribuido al desarrollo de la minería de datos.

Por su parte, Rstudio es una interfaz que permite acceder de forma sencilla a R, en donde se requiere de la instalación previa de R para poder disfrutar de su entorno amigable. Dentro del contexto de este análisis, mencionaremos a continuación las paquetes a utilizar: MASS (Modern Applied Statistic with S), o Estadistica moderna aplicada con S. Fue creado por (Venables, Firth, Bates, Hornik, & Gebhardt, 1998) , y contribuirá con los modelos de regresión logística y Bagging. Adicionalmente utilizaremos los paquetes dependientes de esta librería los cuales son: lattice, nlme, nnet y survival. Por su parte DAAG (Maindonald & Braun, 2003-2017), por sus siglas en español, Análisis y Grafica de datos, ayudará a la elaboración de árboles de decisión y randomForest creado por (Breiman, Cutler, Liaw, & Wiener, s.f.) con Bosques Aleatorios.

En la plataforma web: https://www.r-project.org/index.html, encontrarán más información sobre paquetes o instalación de R y para indagar más sobre Rstudio se recomienda ir a la siguiente página oficial http://www.rstudio.org.

2.8.2 MICROSOFT EXCEL

Excel es una aplicación de "Microsoft Office" para hojas de cálculo que permite crear tablas, cálculos y análisis de datos.[3] Microsoft comercializó inicialmente un programa de hojas de cálculo denominado multiplan en 1982, luego publicó la primera versión de Excel para Mac en 1985 y la primera versión para Windows en 1987; así, con el paso del tiempo ha ido impulsando ventaja competitiva frente a sus competidores lanzando al mercado cada dos años versiones nuevas para Excel. Actualmente cuenta con la versión 2017.

Al igual que R, utilizaremos Excel para la elaboración de las matrices de transición, roll rate, tablas performance del modelo I y II, gráficas y tablas de resumen.

.

[3]Mansfield, Ron (1994) (en español). "*Guía completa para Office de Microsoft"*. traducción Jaime Schlittler. México, D.F.: Ventura.

CAPÍTULO 3

ANÁLISIS DE RESULTADOS

3.1. PROCEDIMIENTO DEL CÁLCULO

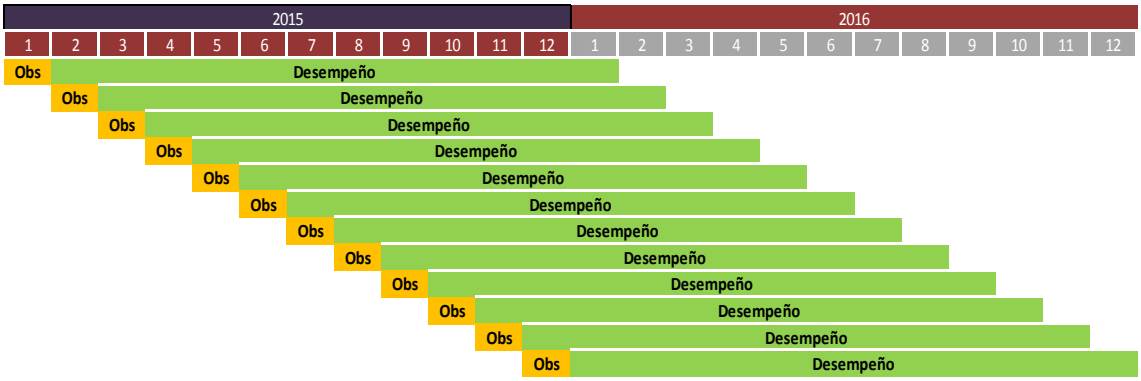
3.1.2 DISEÑO MUESTRAL

De la población global, se seleccionó una muestra aleatoria representativa de 5.444 clientes con año de aprobación y consumo en el periodo de enero a diciembre 2015 de acuerdo a cada uno de los puntos de observación definidos en la ventana de desempeño. (Ver figura 14)

Para los clientes en la muestra, se obtuvo el detalle histórico de variables de desempeño para un horizonte o ventana de análisis de 12 meses de acuerdo a la figura 3, con el objetivo de definir la variable dependiente.

Además para cada observación se obtuvo variables demográficas y de comportamiento las cuales denominaremos variables independientes y se definieran como modelos I y II de acuerdo a las conceptos establecidos.

Figura 14: Ventana de desempeño de análisis



Fuente: Elaboración Propia

3.1.3 DEFINICIÓN DE LA VARIABLE DEPENDIENTE

La definición de default es importante dentro de la construcción de los modelos de scoring o riesgo de crédito, ya que depende de éste obtener la variable que se va a predecir.

Existen dos metodologías para definir el punto de default o el punto donde los clientes caerán en mora: Roll Rate y Matrices de transición.

3.3.2.1 MATRIZ ROLL RATE

Roll rate permite identificar el deterioro crediticio de los clientes, contrastando el tramo de vencimiento al mes de análisis versus el tramo de vencido al mes siguiente dentro del horizonte de tiempo de 12 meses de desempeño establecido.

El objetivo de realizar esta metodología es que se permita definir el punto a partir del cual la tasa de deterioro de los clientes se empeora o estabiliza permitiendo identificar aquellos clientes potenciales a caer en mora dentro de la institución.

A continuación se ilustra la distribución de la tasa de avance por tramos de vencimiento; en donde avanza demuestra el deterioro de la edad de vencido actual al mes siguiente considerando las variables de la figura 3.

Tabla 3: Tasa de Avance con Metodología Roll Rate por Tramos de Vencido

Tramos	Avanza				Total No.	%Total Fila
	Si		No			
	No.Clientes	%Fila	No. Clientes	%Fila		
Sin Vencido		0,00%	3.280	100,00%	3.280	100,0%
De 1-15 días	8	44,44%	10	55,56%	18	100,0%
De 16-30 días	275	55,00%	225	45,00%	500	100,0%
De 31-45 días	3	60,00%	2	40,00%	5	100,00%
De 46-60 días	166	48,26%	178	51,74%	344	100,00%
De 61-75 días	4	66,67%	2	33,33%	6	100,00%
De 76-90 días	127	54,51%	106	45,49%	233	100,00%
De 91-120 días	111	76,55%	34	23,45%	145	100,00%
De 121-150 días	85	81,73%	19	18,27%	104	100,00%
Mayor a 150 días	809	100,00%		0,00%	809	100,00%
Total general	**1.588**	**29,17%**	**3.856**	**70,83%**	**5.444**	**100,0%**

Fuente: Elaboración Propia

Tabla 4: Resumen Tasa de Avance con Metodología Roll Rate

Tramos	Avanza				Total No.	%Total Fila
	Si		No			
	No.Clientes	%Fila	No. Clientes	%Fila		
Sin Vencido		0,00%	3.280	100,00%	3.280	100,0%
De 1-30 días	283	54,63%	235	45,37%	518	100,0%
Mayor a 30 días	1.305	79,28%	341	20,72%	1.646	100,00%
Total general	**1.588**	**29,17%**	**3.856**	**70,83%**	**5.444**	**100,0%**

Fuente: Elaboración Propia

Los criterios establecidos para la definición de desempeño de clientes buenos, indeterminados y malos en base a la tabla 4 son los siguientes: zona verde, aquellos clientes en donde la tasa de avance de mora o deterioro se encuentra por debajo del 42% y que representa a los clientes que se encuentran al día. La zona amarilla define a los clientes indeterminados. La tasa de avance en esta zona se encuentra entre el 42% y 54% y corresponde al tramo entre 1 a 30 días de vencido. Finalmente tenemos a los clientes malos con una tasa de avance superior al 54%, es decir es más probable que un cliente caiga en incumplimiento de pago si su mora es mayor a los 30 días de vencido.

Usando la matriz Roll rate, definimos la variable dependiente en función de los días de mora como:

- **Clientes Buenos:** Aquellos clientes sin atrasos dentro de la ventana de desempeño, es decir 0 días de mora dentro de la ventana de análisis.
- **Clientes Indeterminados:** Clientes que no se definieron como buenos o malos, y serán aquellos que se encuentren en el tramo de 1 a 30 días de vencido dentro de la ventana de desempeño
- **Clientes Malos:** Aquellos clientes con atrasos mayores a 30 días dentro de la ventana de desempeño.

A continuación se muestra en resumen las definiciones de desempeño:

Figura 15: Definiciones de Desempeño

Definición	No. Clientes	%No.
BUENO	3.280	60,2%
INDETERMINADO	518	9,5%
MALO	1.646	30,2%
Total general	**5.444**	**100,0%**

Fuente: Elaboración Propia

3.3.2.2 DEFINICIÓN DE EXCLUSIONES

En la etapa de desarrollo para la elaboración del modelo excluimos a los clientes catalogados como indeterminados debido a que pertenecen a la zona amarilla y resulta complejo definir si son clientes buenos o malos, por ende para evitar el sesgo es recomendable extraerlos al momento de realizar el modelamiento.
El conjunto de datos para el análisis se reduce entonces a 4.926 clientes, en donde el 30.2% son considerados como malos y el 60.2% como buenos.

Tabla 5: Definición de la Variable Dependiente

Definición	No. Clientes	%No.
BUENO	3.280	60,2%
MALO	1.646	30,2%
Total general	**4.926**	**90,5%**

Fuente: Elaboración Propia

3.3.2.3 MUESTREO

El conjunto de datos fue dividido en 2 submuestras: La primera submuestra a la cual denominaremos datos de entrenamiento se utilizó para entrenar el modelo. La segunda submuestra a la cual denominaremos datos de prueba, se la usó para la validación de la técnica con el objetivo de evaluar su capacidad predictiva. Si analizamos la técnica usando los datos de entrenamiento, la evaluación sería injusta, dado que se utilizaría los mismos datos en donde se construye la técnica para su evaluación. Se busca más bien evaluar la capacidad de la técnica en generalizar su capacidad predictiva.

Para entrenar el modelo se consideró el 65% de clientes y para la fase de validación al 35%, de un total de 4.926 clientes.

A continuación se presenta las estadísticas del muestreo aleatorio simple:

Tabla 6: Tabla descriptiva de los datos de Prueba y Entrenamiento

Base	No.Clientes	%No.
Prueba	1.724	35,00%
Entrenamiento	**3.202**	**65,00%**
Total general	**4.926**	**100,00%**

Fuente: Elaboración Propia

Finalmente, la variable dependiente se catalogará como 1 si el cliente cae en default o en incumplimiento de pago; y 0 caso contrario.

Tabla 7: Codificación Variable Dependiente: Morosidad

Condición	Codificación
Moroso	1
No Moroso	0

Fuente: Elaboración Propia

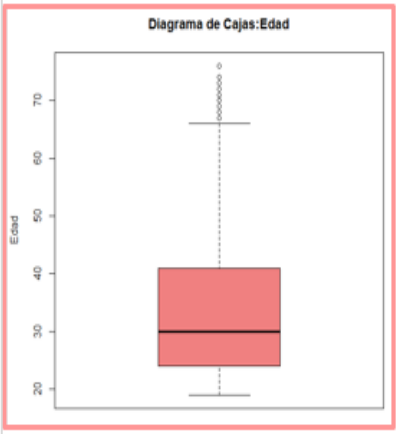
3.3.3 DEFINICION DE LAS VARIABLES INDEPENDIENTES

Las variables independientes permitirán explicar a la variable de estudio.
Para las respectivas variables se realizó un análisis Univariante y Bivariante.

El análisis univariante, permite analizar de forma individual a cada variable e identificar si existen datos faltantes y valores atípicos que afecten al rendimiento del modelo.

La figura 19, por ejemplo, muestra el análisis descriptivo de la variable edad. La edad promedio de los clientes es de 30 años, la mínima de 19 y la máxima de 76 años respectivamente, mientras que menos del 5% de los clientes poseen edades entre los 19 y 21 años.

Figura 16: Análisis Univariado de la Variable Edad



Descriptivos	Estadísticos
Media	32,92
Mediana	30
Varianza	162,91
Desv. Tipica	12,76
Min	19
Max	76
Asimetría	1,015158
Curtosis	3,098374

	Percentiles						
Edad	5	10	25	50	75	90	95
	21	22	24	30	41	54	61

Fuente: Elaboración Propia

Por su parte el análisis Bivariante permite analizar dos variables de manera simultánea.

En la tabla 8, se puede observar que la variable edad se encuentra correlacionada con la variable morosidad, dado que a menor edad existe mayor probabilidad de que el cliente caiga en mora según las tasas de deterioro.

Para confirmar la correlación de ambas variables utilizaremos el test Chi-cuadrado para corroborar la dependencia o independencia de las mismas en donde en efecto de acuerdo al valor p obtenido 3.083e-08, (el cual es cercano a 0), podemos mencionar que existe evidencia estadística para rechazar la hipostasis nula de que las variables de edad y default son independientes.

Tabla 8: Análisis Bivariado de la Variable Edad vs Morosidad

INDEPENDIENTE	BUENOS		MALOS		TOTAL		%DIST	%ACUM
EDAD	No.	%	No.	%	No.	%	TOTAL	TOTAL
19-24	650	61,85%	401	38,15%	1051	100,00%	21,34%	21,34%
24-29	755	65,31%	401	34,69%	1156	100,00%	23,47%	44,80%
29-34	422	64,13%	236	35,87%	658	100,00%	13,36%	58,16%
>34	1453	70,50%	608	29,50%	2061	100,00%	41,84%	100,00%
Total general	**3280**	**66,59%**	**1646**	**33,41%**	**4926**	**100,00%**	**100,00%**	

```
> prueba

        Pearson's Chi-squared test

data:  tabla8
X-squared = 59.256, df = 12, p-value = 3.083e-08
```

Fuente: Elaboración Propia

Una vez realizados los respectivos análisis univaridos y bivariados se procederá a realizar las respectivas codificaciones para las variables independientes para los modelos I y II.

3.3.3.1 CODIFICACIÓN DE LAS VARIABLES INDEPENDIENTES PARA MODELO I

A continuación se describe la codificación de las variables independientes que ingresaran al modelo I. En este bloque solo se consideran las variables demográficas de los clientes:

- **Estado civil:** Indica el estado civil registrado por la institución.

Tabla 9: Codificación Variable Independiente: Estado Civil

Condición	Codificación
Soltero	1
Otros	0

Fuente: Elaboración Propia

- **Género:** Indica el género registrado por la institución.

Tabla 10 Codificación Variable Independiente: Género

Condición	Codificación
Femenino	1
Masculino	0

Fuente: Elaboración Propia

Formalidad: Indica la formalidad laboral del cliente, en donde alta se refiere a si un cliente se encuentra en relación de dependencia, es decir si está asegurado o posee negocio propio con legalidad de RUC, mientras que Otros, indica si el cliente se encuentra en relación de dependencia o si realiza alguna actividad informal.

Tabla 11: Codificación Variable Independiente: Formalidad

Condición	Codificación
Alta	1
Otros	0

Fuente: Elaboración Propia

- **Edad:** Edad del cliente.

- **Tiene teléfono convencional:** Indica si posee teléfono convencional o no registrado por la institución.

Tabla 12 Codificación Variable Independiente: Tiene teléfono convencional

Condición	Codificación
Tiene teléfono convencional	1
Caso Contrario	0

Fuente: Elaboración Propia

- **Tiene teléfono celular:** Indica si posee teléfono celular o no registrado por la institución.

Tabla 13: Codificación Variable Independiente: Tiene teléfono celular

Condición	Codificación
Tiene teléfono celular	1
Caso Contrario	0

Fuente: Elaboración Propia

- **Tipo de propiedad:** Indica si la vivienda del cliente es propia. Mientras que "Otros" aplica cuando la vivienda es alquilada, de padres, familiar etc.

Tabla 14: Codificación Variable Independiente: Tipo de Propiedad

Condición	Codificación
Propia	1
Otros	0

Fuente: Elaboración Propia

- **Tipo de Construcción:** Indica si el tipo de construcción de la vivienda del cliente "cemento", mientras que "Otros" aplica si el tipo de la vivienda es de caña, de madera, etc.

Tabla 15: Codificación Variable Independiente: Tipo de Construcción

Condición	Codificación
Cemento	1
Otros	0

Fuente: Elaboración Propia

- **Tiempo de estabilidad domiciliaria:** Indica el tiempo de estabilidad domiciliaria en meses del cliente registrado por la institución.

- **Ingresos:** Indica el ingreso del cliente en dólares registrado por la institución.

- **Provincia de Origen:** Indica la provincia de origen del cliente registrado por la institución. (Por motivos de sigilo comercial no se menciona las ciudades, solo sus códigos).

Tabla 16: Codificación Variable Independiente: Provincia de Origen

Condición	Codificación
CHI,ESM,RS,MO,PZA,PA,ZC,O,STO	1
Otros	0

Fuente: Elaboración Propia

- **Región de consumo:** Indica la región de consumo de los clientes registrados por la institución.

Tabla 17: Codificación Variable Independiente: Región de Consumo

Condición	Codificación
Costa - Centro	1
Otros	0

Fuente: Elaboración Propia

- **Ciudad de consumo:** Indica la ciudad de consumo del cliente registrado por la institución. (Por motivos de sigilo comercial no se menciona las ciudades, solo sus códigos).

Tabla 18: Codificación Variable Independiente: Consumo de Consumo

Condición	Codificación
TRN, QO, QD,SO,PO	1
Otros	0

Fuente: Elaboración Propia

- **Agencia de consumo:** Indica la ciudad de consumo del cliente registrado por la institución. (Por motivos de sigilo comercial no se menciona las agencias, solo sus códigos).

Tabla 19: Codificación Variable Independiente: Agencia de Consumo

Condición	Codificación
TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QD-ALM, QO-ALM5, QO-ALM8, SO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4	1
Otros	0

Fuente: Elaboración Propia

- **Fuente de Ingreso:** Indica la fuente de ingreso del cliente es laboral caso contrario negocio.

Tabla 20: Codificación Variable Independiente: Fuente de Ingreso

Condición	Codificación
Laboral/Relación de dependencia	1
Otros	0

Fuente: Elaboración Propia

- **Profesión:** Indica la profesión del cliente.

Tabla 21: Codificación Variable Independiente: Profesión

Condición	Codificación
Jornalero **Comerciante** **Vendedor** **Mecánico** **Estudiante** **Empleado Particular** **Agricultor** **Negociante** **Ayudante** **Albañil** **Empleada Doméstica** **Obrero** **Soldador** **Empleado Público** **Electricista** **Auxiliar** **Taxista** **Peluquero** **Recaudador** **Economista** **Supervisor** **Secretaria** **Acuicultor** **Camarero** **Doctor** **Militar** **Lotero** **Bombero** **Zapatero** **Apoderado** **Constructor** **Prensista**	1
Otros	0

Fuente: Elaboración Propia

- **Avalúo activo del vehículo:** Indica el valor en dólares del avalúo vehículo del cliente.

- **Profesión cónyuge:** Indica la profesión del cónyuge del cliente.

Tabla 22: Codificación Variable Independiente: Profesión de Cónyuge

Condición	Codificación
Que haceres Domésticos **Profesor** **Policía** **Panadero** **Peluquero** **Tecnólogo** **Soldador** **Pescador** **Tejedor** **Vendedor** **Plomero** **Pintor**	1
Otros	0

Fuente: Elaboración Propia

- **Salario cónyuge:** Indica el salario del cónyuge del cliente en dólares

- **Avalúo activo del vehículo del cónyuge:** Indica el valor en dólares del avalúo del vehículo del cónyuge del cliente.

3.3.3.2 CODIFICACIÓN DE LAS VARIABLES INDEPENDIENTES PARA MODELO II

El segundo modelo de este estudio considera variables demográficas y de comportamiento. A continuación la codificación se detalla la codificación usada:

- **Estado actual de la tarjeta:** Indica si la tarjeta institucional del cliente se encuentra actualmente bloqueada o no.

Tabla 23: Codificación Variable Independiente: Estado actual de la tarjeta

Condición	*Codificación*
Bloqueada	1
Otros	0

Fuente: Elaboración Propia

Motivo del bloqueo de la tarjeta: Indica los motivos por los cuales la tarjeta del cliente se encuentra actualmente bloqueada por Dación si al cliente se le retiró el bien por falta de compromiso de pago. Reestructuración si al cliente se le realizó un nuevo financiamiento por atrasos de pago con la institución, entre otros.

Tabla 24: Codificación Variable Independiente: Motivo de Bloqueo de la Tarjeta

Condición	Codificación
Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración	1
Otros	0

Fuente: Elaboración Propia

Venta de cartera, Morosidad y Black Friday: Estados de clientes que superan los días de mora permitidos por la institución.

- **Cupo mensual disponible actual:** Es el cupo mensual disponible actual en dólares que el cliente registra dentro de la institución.

- **Cupo mensual facturado:** Es el cupo mensual facturado en dólares de acuerdo al plazo de crédito pactado con el cliente y la institución.

- **Cupo mensual asignado actual:** Es el cupo mensual asignado actual en dólares que el cliente posee dentro de la institución.

- **Cupo mensual aprobado inicial:** Es el cupo mensual aprobado inicialmente al momento en el que se le otorgo el crédito.

- **Número de Aumentos de cupo:** Es el número de aumentos de cupos efectivos que se le realizo al cliente dentro del periodo de análisis.

- **Número de negociaciones de cupo:** Es el número de negociones de aumentos de cupo que el cliente ha realizado dentro del periodo de análisis.

- **%Extracupo:** Es la razón porcentual entre el cupo mensual disponible actual y el cupo mensual asignado actual.

3.2 PRESENTACIÓN DE RESULTADOS

3.2.1 REGRESIÓN LOGÍSTICA

3.2.1.1 TABLA PERFORMACE PARA MODELO I

3.2.1.1.1 TABLA PERFORMANCE PARA SUBMUESTRA DE ENTRENAMIENTO:

De acuerdo a las definiciones anteriormente planteadas, A continuación se detalla la tabla performance o de modelamiento para la submuestra de entrenamiento del modelo I considerando solo las variables demográficas de los sujetos a crédito:

Tabla 25: Tabla Scorecard para Modelamiento de Entrenamiento del Modelo I

KS	19%	ROC	68%	GINI	26%

Score	Total			Malos			Default	
	No.	%No.	%Acum	No.	%No.	%Acum	%Malos	%Decum_Malos
[>=958]	174	5%	5%	28	3%	3%	16%	33%
[944-957]	153	5%	10%	27	3%	5%	18%	34%
[934-943]	169	5%	15%	39	4%	9%	23%	35%
[926-933]	160	5%	20%	33	3%	12%	21%	35%
[918-925]	163	5%	26%	45	4%	16%	28%	36%
[910-917]	157	5%	30%	48	5%	21%	31%	37%
[903-909]	151	5%	35%	45	4%	25%	30%	37%
[895-902]	163	5%	40%	49	5%	30%	30%	38%
[886-894]	170	5%	46%	50	5%	35%	29%	38%
[878-885]	160	5%	51%	44	4%	39%	28%	39%
[870-877]	156	5%	55%	40	4%	43%	26%	40%
[860-869]	157	5%	60%	55	5%	48%	35%	42%
[849-859]	151	5%	65%	50	5%	53%	33%	43%
[834-848]	172	5%	70%	66	6%	59%	38%	44%
[819-833]	150	5%	75%	45	4%	63%	30%	45%
[802-818]	156	5%	80%	69	7%	70%	44%	48%
[774-801]	162	5%	85%	62	6%	76%	38%	49%
[747-773]	163	5%	90%	85	8%	84%	52%	53%
[<=746]	315	10%	100%	168	16%	100%	53%	53%
Total general	3.202	100%		1.048	100%		33%	

Fuente: Elaboración Propia

Dónde:

- **Score:** Es el puntaje de calificación crediticia que recibiría el cliente dependiendo de su comportamiento de pago, la cual se encuentra expresado en intervalos reflejando el valor mínimo y máximo para cada tramo de score.
- **No. Total:** Es el número global de clientes para cada rango de score.
- **%No. Total:** Es la frecuencia global de clientes para cada rango de score.
- **%Acum. Total:** Es la frecuencia acumulada global de clientes para cada rango de score.
- **No. Malos:** Es el número de clientes malos para cada rango de score.
- **%No. Malos**: La frecuencia de clientes malos para cada rangos de score.
- **%Acum. Malos:** Es la frecuencia acumulada de clientes malos para cada rango de score.
- **%Malos Default:** Es la tasa de clientes malos respecto al global de clientes, es decir la probabilidad que alcance el default en una mora mayor o peor a 30 días en los próximos 12 meses. La definición de los días de mora como punto de incumpliendo se definieron la sección 3.3.2.1
- **%Decum Default:** Es la probabilidad máxima des acumulada de clientes malos.

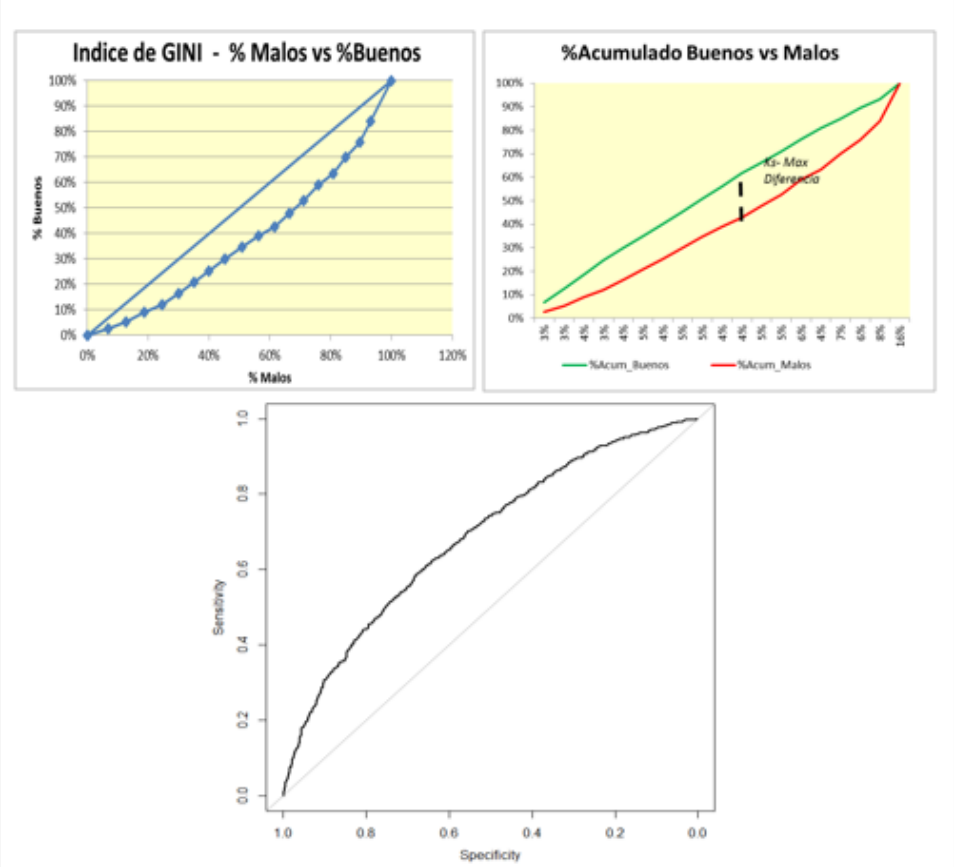
La Tabla 25, muestra la submuestra de modelamiento distribuida en ventiles, el segmento de default es el más importante debido a que muestra la probabilidad de que un cliente caiga en una mora mayor a 30 días de vencido en los siguientes 12 meses. Adicional se ilustran los estadísticos KS: 19%, ROC: 68% y GINI: 26%, siendo estas tasas considerables para un modelo de originación (Modelo I) de acuerdo a los criterios establecidos en las medidas de calidad en la sección 2.2.4.

Es considerable pretender que los estadísticos en mención sean iguales al 100%, pero no sucede en la vida real, dado que todo modelo predictivo maneja su respectivo margen de error. Lo ideal es obtener el modelo con el menor error posible y que sea posible de automatizar para optimizar el otorgamiento de crédito.

Para la interpretación de la tabla performance scorecard, tomaremos como referencia un cliente que sea calificado o puntuado con score 959, éste se encontraría en el 5% de la muestra de desarrollo de acuerdo al porcentaje de clientes acumulados y a su vez presentaría una probabilidad de alcanzar una mora mayor a 30 días en los próximos 12 meses del 16%. Finalmente de acuerdo a la des acumulada los clientes que se estaría dejando fuera tendrían una tasa del 33%.

A continuación se muestran las gráficas GINI, K-S y la curva ROC para la submuestra de desarrollo del modelo I:

Figura 17: GINI,K-S y ROC para Modelamiento de Desarrollo del Modelo I




```
> g$auc
ROC: Area under the curve: 0.684
```

Fuente: Elaboración Propia

La graficas GINI y K-S y ROC representan gráficamente la separación de clientes buenos.

ROC ,representa la sensibilidad y especificidad, en donde la sensibilidad es la probabilidad de que a un cliente bueno la prueba le dé resultado positivo, mientras que especificidad indica la probabilidad de que a un cliente malo la prueba le dé resultado negativo (matriz de confusión).

3.2.1.1.2 TABLA PERFORMANCE PARA SUBMUESTRA DE PRUEBA

Para el contraste del modelamiento de desarrollo se aplicará la misma metodología para la submuestra de prueba, en donde se detalla en la siguiente tabla de resumen scorecard:

Tabla 26: Tabla Scorecard para Modelamiento de Validación del Modelo I

KS	22%	ROC	69%	GINI	28%

Score	Total			Malos			Default	
	No.	%No.	%Acum	No.	%No.	%Acum	%Malos	%Decum_Malos
[>=961]	92	5%	5%	15	3%	3%	16%	35%
[949-960]	83	5%	10%	15	3%	5%	18%	36%
[939-948]	85	5%	15%	20	3%	8%	24%	37%
[929-938]	95	6%	21%	22	4%	12%	23%	37%
[922-928]	85	5%	26%	25	4%	16%	29%	38%
[913-921]	82	5%	30%	17	3%	19%	21%	39%
[906-912]	83	5%	35%	21	4%	23%	25%	40%
[897-905]	88	5%	40%	23	4%	26%	26%	41%
[890-896]	99	6%	46%	32	5%	32%	32%	43%
[884-889]	76	4%	50%	27	5%	36%	36%	44%
[875-883]	86	5%	55%	33	6%	42%	38%	45%
[867-874]	82	5%	60%	35	6%	48%	43%	45%
[855-866]	85	5%	65%	32	5%	53%	38%	45%
[844-854]	89	5%	70%	38	6%	59%	43%	47%
[829-843]	84	5%	75%	32	5%	65%	38%	47%
[810-828]	89	5%	80%	36	6%	71%	40%	49%
[791-809]	83	5%	85%	39	7%	77%	47%	51%
[759-790]	86	5%	90%	48	8%	85%	56%	53%
[<=758]	172	10%	100%	88	15%	100%	51%	51%
Total general	1.724	100%		598	100%		35%	

Fuente: Elaboración Propia

Se puede observar que los resultados de validación el KS(22%), ROC(69%) y GINI (28%) se aproximan al valor obtenido en el modelo I de desarrollo, lo que indica que el modelo no se encuentra sobreajustado.

3.2.1.2 VARIABLES RELEVANTES PARA MODELO I

Para la toma de decisiones del mejor modelo se utilizó la metodología AIC (Criterio de Información de Akaike), el cual relaciona el coeficiente de la verosimilitud con el número de parámetros del modelo.

Dentro de la metodología AIC se utilizó el procedimiento de selección de variables "backward", el cual parte inicialmente con todas las variables descartando paso a paso las variables menos significativas del modelo en donde con la segregación de variables de la última etapa se obtiene el modelo con el menor AIC para definición del modelo final.

Finalmente se obtienen las variables independientes que explican a la variable resultante del modelo:

Figura 18: Resultados del Modelo I de Desarrollo en Regresión Logística

```
                          Estimate   Std. Error  z value  Pr(>|z|)
                            (β̂)        (σβ̂)
(Intercept)              1.1945929   0.3464358   3.448  0.000564 ***
Estado_Civil1            0.3002628   0.1005610   2.986  0.002828 **
Sexo1                   -0.4601487   0.0920760  -4.997  5.81e-07 ***
Edad                    -0.0139366   0.0038717  -3.600  0.000319 ***
Formalidad1             -0.7429720   0.0968630  -7.670  1.72e-14 ***
tienetelefonoConv1      -0.3967228   0.0934103  -4.247  2.17e-05 ***
tienetelefonoCelular1   -0.6718809   0.2727194  -2.464  0.013754 *
Propiedad1              -0.3775365   0.1076685  -3.506  0.000454 ***
TiempoEstabiulidadDom   -0.0014181   0.0003851  -3.682  0.000231 ***
Ingresos_.              -0.0002345   0.0001275  -1.838  0.066004 .
Provincia_Origen1        0.2034809   0.0908104   2.241  0.025044 *
Agencia_consumo1         0.4097805   0.0816142   5.021  5.14e-07 ***
fuente_ingreso          -0.1904067   0.0959675  -1.984  0.047248 *
Avaluo_Act_Vehiculo_.   -0.0001260   0.0001065  -1.183  0.236949
Salario_Conyuge_.       -0.0011269   0.0005622  -2.005  0.045007 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4048.9  on 3201  degrees of freedom
Residual deviance: 3740.3  on 3187  degrees of freedom
AIC: 3770.3

Number of Fisher Scoring iterations: 4
```

Fuente: Elaboración Propia

Según la figura 24, se presentan los siguientes resultados:

- $\widehat{\boldsymbol{\beta}}$: Son los coeficientes estimados asociados a cada una de las variables independientes.

Los coeficientes $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ son utilizados para determinar la probabilidad de que un cliente caiga en default o incurra en morosidad, es decir, la probabilidad de que y=1.

$$\boldsymbol{P}(\boldsymbol{y}=\mathbf{1}|\,\mathbf{x})=\frac{\boldsymbol{e}^{\boldsymbol{z}}}{\mathbf{1}+\boldsymbol{e}^{\boldsymbol{z}}}=\frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{z}}}$$

- Std. Error $\boldsymbol{\sigma}_{\widehat{\boldsymbol{\beta}}}$: Es el error estándar para cada estimación de los coeficientes de $\widehat{\boldsymbol{\beta}}$.
- Z Value: Estadístico de prueba Z que evalúa la diferencia entre un estadístico observado y su parámetro poblacional en unidades de desviación estándar.
- Pr(>|**z**|): Es el Valor p del estadístico de Prueba Z y representa el nivel de significancia más pequeño que conduce al rechazo de la hipótesis nula.

De acuerdo a la interpretación de los signos de los coeficientes de $\widehat{\boldsymbol{\beta}}$ se tiene que los clientes con estado civil soltero, incrementan la probabilidad de ser morosos al igual que la variables provincia de origen y agencias de consumo, mientras que las variables Sexo, Edad, Formalidad, Posee teléfono convencional, Posee teléfono celular, Propiedad, Tiempo de estabilidad domiciliaria y salario del cónyuge disminuyen la probabilidad de que un cliente sea moroso de acuerdo a la codificación obtenida en la Tabla 20.

Por motivos de sigilo de información, no se puede transcribir la ecuación del modelo, sin embargo recordarnos que la ecuación del modelo logístico se define como:

$$z=Logit\ (P)=ln\left(\frac{p}{1-p}\right)=\boldsymbol{\beta_0}+\boldsymbol{\beta_1}x_1+\boldsymbol{\beta_2}x_2+\cdots+\boldsymbol{\beta_q}x_q$$

Tabla 27: Interpretación de Variables en base a los signos de los coeficientes de $\widehat{\beta}$ del Modelo I de Entrenamiento

Variables	Descripción	Códigos	Signo	Explicación
Estado_Civil	Soltero	1	Positivo	Clientes con estado civil soltero incrementan la probabilidad de ser morosos
	Otros	0		
Sexo	Mujer	1	Negativo	Clientes mujeres dismuyen la probabilidad de ser morosos
	Hombre	0		
Edad	Numérica	-	Negativo	La edad disminuye la probabilidad de ser morosos
Formalidad	Alta	1	Negativo	Clientes formales disminuyen la probabilidad de ser morsosos
	Otros	0	Negativo	
Tiene Teléfono convencional	Posee Teléfono convencional	1	Negativo	Clientes que poseen teléfono convencional dismuyen la probabilidad de ser morosos
	No posee Teléfono convencional	0	Negativo	
Tiene Teléfono celular	Posee Teléfono celular	1	Negativo	Clientes que poseen teléfono celular dismuyen la probabilidad de ser morosos
	No posee teléfono celular	0	Negativo	
Propiedad	Propia	1	Negativo	Clientes que viven en vivienda propia disminuyen la probabildad de ser morosos
	Otros	0	Negativo	
Tiempo de Estabilidad domiciliaria	Numérica	-	Negativo	El tiempo de estabilidad domiciliaria disminuye la probabilidad de ser morosos
Provincia de Origen	CHI,ESM,RS,MO,PZA,PA,ZC,O,SO	1	Positivo	Las provincias codificadas descritas incrementan la probabilidad de ser morosos
	Otros	0		
Agencia de consumo	TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QO-ALM, QO-ALM5, QO-ALM8, STO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4	1	Positivo	Las agencias codificadas descritas incrementan la probabilidad de ser morosos
	Otros	0		
Fuente de Ingreso	Laboral/Relación de dependencia	1	Negativo	La fuente de ingreso laboral disminuye la probabilidad de ser moroso
	Otros	0		
Avalúo del Activo del Vehículo	Numérica	-	Negativo	El avalúo del activo del vehículo disminuye la probabilidad de ser morosos
Salario Conyuge	Numérica	-	Negativo	El salario del conyuge disminuye la probabilidad de ser morosos

Fuente: Elaboración Propia

Analizando el valor p, se puede evidenciar que las variables mayormente significativas son Sexo, Edad, Formalidad, Tiene teléfono convencional, propiedad, tiempo de estabilidad domiciliaria, y agencias de consumo dado que el valor p es menor a 0.001, y existen dos variables que son menos significativas como: avalúo del activo del vehículo e ingresos percibidos por lo tanto existe evidencia estadística para no rechazar la hipótesis nula de que los betas sean iguales a cero. Mientras que el valor p para el resto de variables es inferior a 0.05 por lo tanto existe evidencia estadística para rechazar la hipótesis nula de que los $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ sean iguales a 0 lo cual indica que existe relación entre la variable dependiente y las covariables o variables explicativas.

En este caso no se descartarán las variables no significativas dado que el valor p es un valor referencial y no implica que al extraerlas el error de predicción disminuya por ende se considerará las variables finalmente obtenidas como último paso del AIC.

3.2.1.2.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I PARA REGRESIÓN LOGISTICA

Respecto a la matriz de confusión se procedió a clasificar a los clientes o sujetos de crédito utilizando como umbral óptimo $\boldsymbol{c}$=0.48 con el objetivo de evitar el incremento de falsos positivos (Valor de z extremadamente <= 0.48) o negativos (Valor de z extremadamente >0.48) dentro de la clasificación, por ende si el valor obtenido al reemplazar cada uno de los factores de la ecuación es menor que 0.48 se clasificará como no moroso, caso contrario se clasificará como moroso. De acuerdo a la tabla 28 se obtiene que, el 89.97% de los clientes no morosos fueron clasificados correctamente, mientras que el 10.03% fueron clasificados como morosos; y referente a los clientes morosos el 30.78% fueron clasificados correctamente y los restantes se encasillaron como no morosos.

Además se obtiene la precisión de predicción y de error que corresponden al 70.58% y 29.42% respectivamente en donde se justifica dicho nivel de error debido a las características de las variables que tiene el modelo.

Tabla 28: Matriz de Confusión de Regresión Logística para Modelo I de Desarrollo

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	1938	216	**2154**
Moroso (y=1)	726	322	**1048**
Total	**2664**	**538**	**3202**

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	**89,97%**	10,03%	**100,00%**
Moroso (y=1)	69,27%	**30,73%**	**100,00%**

```
> precision3 <- sum(diag(confusion3))/sum(confusion3)
> precision3
[1] 0.7058089
> error.prediction3 <- 1-precision3
> error.prediction3
[1] 0.2941911
```

Fuente: Elaboración Propia

Finalmente se contrasta la matriz de confusión de desarrollo y de validación para evidenciar las tasas de precisión de la clasificación y error de predicción:

Tabla 29: Matriz de Confusión de Regresión Logística para Modelo I de Validación

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	981	145	**1126**
Moroso (y=1)	384	214	**598**
Total	**1365**	**359**	**1724**

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	**87,12%**	12,88%	**100,00%**
Moroso (y=1)	64,21%	**35,79%**	**100,00%**

```
> precision3 <- sum(diag(confusion3))/sum(confusion3)
> precision3
[1] 0.6931555
> error.prediction3 <- 1-precision3
> error.prediction3
[1] 0.3068445
g<-roc(Datos2$Default~prediccion3)
> g$auc
Area under the curve: 0.6938
```

Fuente: Elaboración Propia

3.2.1.3 TABLA PERFORMACE PARA MODELO II

3.2.1.3.1 TABLA PERFORMANCE PARA SUBMUESTRA DE ENTRENAMIENTO:

Para el desarrollo del Modelo II, se toma en consideración las variables demográficas y de comportamiento de los clientes sujetos a crédito en donde a continuación se presenta la tabla performance para la submuestra de entrenamiento:

Tabla 30: Tabla Scorecard para Modelamiento de Desarrollo del Modelo II

KS	74%	ROC	94%	GINI	83%

Score	Total			Malos			Default	
	No.	%No.	%Acum	No.	%No.	%Acum	%Malos	%Decum_Malos
[>=996]	169	5%	5%	1	0%	0%	1%	33%
[991-995]	179	6%	11%	3	0%	0%	2%	35%
[987-990]	137	4%	15%	1	0%	0%	1%	37%
[981-986]	162	5%	20%	4	0%	1%	2%	38%
[974-980]	162	5%	25%	3	0%	1%	2%	41%
[964-973]	169	5%	31%	3	0%	1%	2%	43%
[951-963]	148	5%	35%	7	1%	2%	5%	46%
[935-950]	162	5%	40%	9	1%	3%	6%	49%
[913-934]	156	5%	45%	9	1%	4%	6%	53%
[885-912]	160	5%	50%	6	1%	4%	4%	57%
[846-884]	158	5%	55%	25	2%	7%	16%	63%
[781-845]	**160**	**5%**	**60%**	**34**	**3%**	**10%**	21%	68%
[703-780]	160	5%	65%	79	8%	18%	49%	74%
[628-702]	162	5%	70%	115	11%	29%	71%	77%
[546-627]	159	5%	75%	118	11%	40%	74%	78%
[435-545]	159	5%	80%	120	11%	51%	75%	79%
[321-434]	160	5%	85%	113	11%	62%	71%	80%
[220-320]	161	5%	**90%**	113	11%	**73%**	70%	83%
[<=219]	319	10%	**100%**	285	27%	**100%**	89%	89%
Total general	**3.202**	**100%**		**1.048**	**100%**		**33%**	

Fuente: Elaboración Propia

La tabla 30, ilustra a la muestra de modelamiento distribuida en ventiles, el segmento de default para la definición del segundo modelo sigue siendo el más importante debido a que muestra la probabilidad de que un cliente caiga en una mora mayor a 30 días de vencido en los siguientes 12 meses de desempeño.

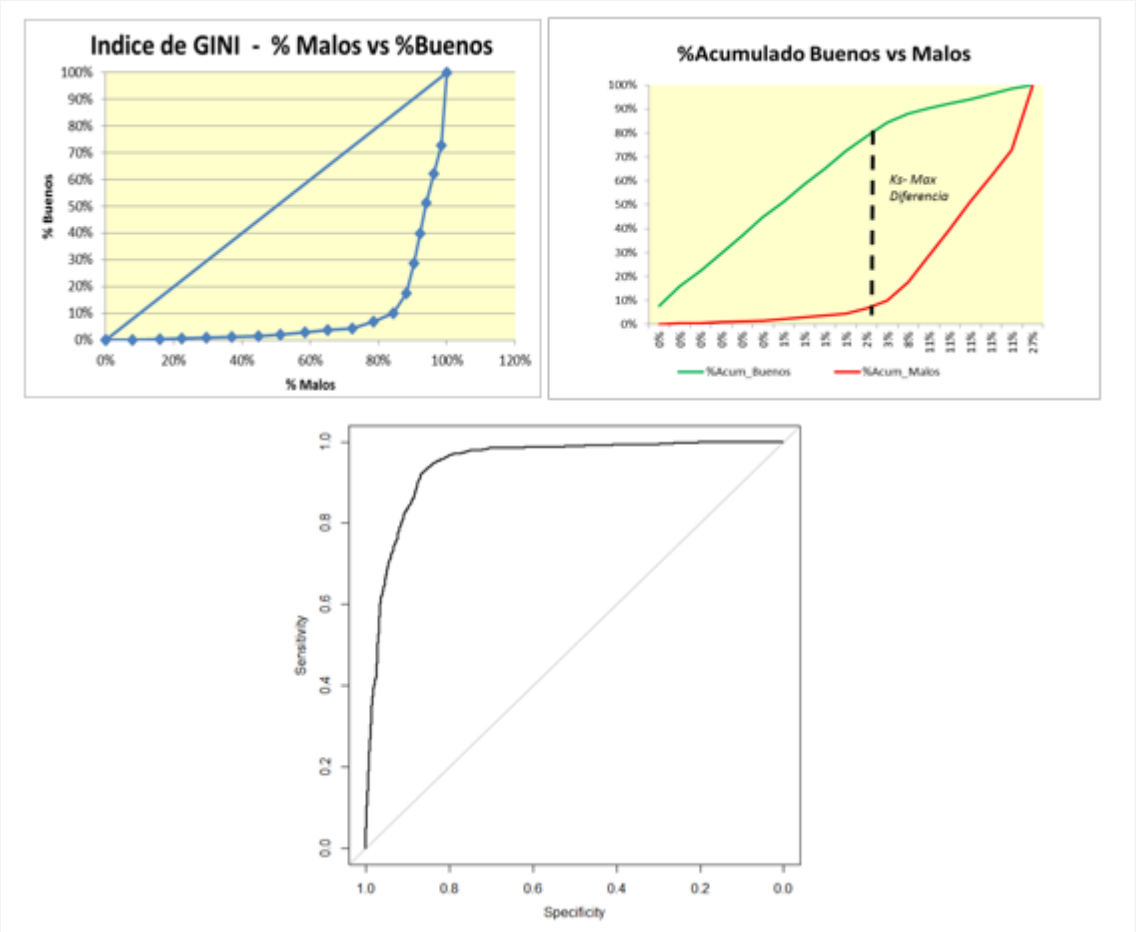
Adicionalmente se muestran los estadísticos KS: 74%, ROC: 94% y GINI: 83%, siendo estas tasas considerables para un modelo de comportamiento (modelo II) de acuerdo a los criterios establecidos en las medidas de calidad.

Para la interpretación de la tabla performance del modelo II, se considera como referencia un cliente que sea calificado o puntuado con score 999, éste se encontraría en el 5% de la muestra de desarrollo de acuerdo al porcentaje de clientes acumulados y a su vez presentaría una probabilidad alcanzar una mora mayor a 30 días en los próximos 12 meses del 1%.

Finalmente de acuerdo a la desacumulada, los clientes que se estaría dejando fuera tendrían una tasa del 33%.

A continuación se muestran las gráficas GINI, K-S y la curva ROC para la submuestra de desarrollo del modelo II:

Figura 19 :Curva ROC para Modelamiento de Desarrollo del Modelo II




```
> g$auc
ROC: Area under the curve: 0.9448
```

Fuente: Elaboración Propia

Las gráficas GINI, K-S y ROC representan la separación de clientes buenos y malos.

ROC, representa la sensibilidad y especificidad, en donde la sensibilidad es la probabilidad de que a un cliente bueno la prueba le dé resultado positivo, mientras que especificidad indica la probabilidad de que a un cliente malo la prueba le dé resultado negativo (matriz de confusión).

Si se compara las gráficas de calidad del modelo I vs el modelo II, se puede evidenciar que existe mayor poder de discriminación en el modelo II, debido a que consideramos variables demográficas más variables de comportamiento de pago lo que hace que el modelo sea más robusto.

3.2.1.3.2 TABLA PERFORMANCE PARA SUBMUESTRA DE PRUEBA:

Los resultados para la muestra de prueba del modelo II se detallan a continuación:

Tabla 31: Tabla Scorecard para Modelamiento de Prueba del Modelo II

KS	77%	ROC	93%	GINI	86%

Score	Total			Malos			Default	
	No.	%No.	%Acum	No.	%No.	%Acum	%Malos	%Decum_Malos
[>=998]	103	6%	6%	1	0%	0%	1%	35%
[996-997]	75	4%	10%	-	0%	0%	0%	37%
[993-995]	95	6%	16%	2	0%	1%	2%	39%
[990-992]	107	6%	22%	-	0%	1%	0%	41%
[988-989]	63	4%	26%	1	0%	1%	2%	44%
[985-987]	84	5%	31%	3	1%	1%	4%	46%
[981-984]	82	5%	35%	1	0%	1%	1%	49%
[974-980]	93	5%	41%	6	1%	2%	6%	53%
[965-973]	77	4%	45%	4	1%	3%	5%	57%
[932-964]	84	5%	50%	4	1%	4%	5%	61%
[820-931]	85	5%	55%	11	2%	6%	13%	67%
[521-819]	88	5%	60%	26	4%	10%	30%	73%
[374-520]	86	5%	65%	49	8%	18%	57%	78%
[293-373]	87	5%	70%	61	10%	28%	70%	81%
[245-292]	87	5%	75%	62	10%	39%	71%	83%
[205-244]	87	5%	80%	67	11%	50%	77%	86%
[174-204]	83	5%	85%	71	12%	62%	86%	88%
[139-173]	86	5%	90%	79	13%	75%	92%	89%
[<=138]	172	10%	100%	150	25%	100%	87%	87%
Total general	1.724	100%		598	100%		35%	

Fuente: Elaboración Propia

En dónde para los resultados de validación el KS (74%), ROC (94%) y GINI (83%) se aproximan a las tasas del modelo II de desarrollo.

3.2.1.4 VARIABLES RELEVANTES PARA MODELO II

Para la toma de decisiones del mejor modelo se utilizó la metodología AIC (Criterio de Información de AKaike), en donde finalmente se obtienen las variables independientes que explican a la variable resultante del modelo:

Figura 20: Resultados del Modelo II de Desarrollo en Regresión Logística

	Estimate $(\widehat{\boldsymbol{\beta}})$	Std. Error $(\sigma_{\widehat{\beta}})$	z value	Pr(>|z|)	
(Intercept)	-5.452e-01	5.999e-01	-0.909	0.36349	
Sexo1	-5.508e-01	1.368e-01	-4.025	5.69e-05	***
Edad	-8.913e-03	5.298e-03	-1.682	0.09249	.
Formalidad1	-3.048e-01	1.443e-01	-2.113	0.03462	*
tieneTelefonoConv1	-2.343e-01	1.403e-01	-1.670	0.09500	.
tieneTelefonoCelular1	-1.034e+00	4.350e-01	-2.377	0.01745	*
TiempoEstabiulidadDom	-1.744e-03	5.590e-04	-3.120	0.00181	**
Provincia_Origen1	3.907e-01	1.439e-01	2.714	0.00664	**
Agencia_consumo1	3.541e-01	1.270e-01	2.789	0.00529	**
estado_tc_actual1	1.904e+00	2.908e-01	6.546	5.90e-11	***
motivo_bloqueo_actual1	2.982e+00	2.230e-01	13.376	< 2e-16	***
cupo_mensual_disponible_actual_.	-4.066e-02	4.104e-03	-9.909	< 2e-16	***
cuota_mensual_facturada_.	-3.517e-02	3.428e-03	-10.258	< 2e-16	***
cupo_mensual_aprobado_inicial_.	1.522e-02	4.950e-03	3.074	0.00211	**
No_aumentos_cupo	5.745e-01	1.406e-01	4.085	4.40e-05	***
X.Extracupo	3.169e-04	5.381e-05	5.890	3.86e-09	***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4048.9  on 3201  degrees of freedom
Residual deviance: 1751.0  on 3186  degrees of freedom
AIC: 1783

Number of Fisher Scoring iterations: 7
```

Fuente: Elaboración Propia

De acuerdo a la interpretación de los signos de los coeficientes de $\widehat{\boldsymbol{\beta}}$ se tiene que los clientes con sexo catalogadas como mujer dismuyen la probabilidad de ser morosos al igual que edad, formalidad, tiene teléfono celular, tiempo de estabilidad domiciliaria, cupo mensual disponible actual y cuota mensual facturada, mientras que provincia de origen, agencia de consumo, estado actual de la tarjeta, motivo del bloqueo de la tarjeta, cupo mensual asignado actual, cupo mensual aprobado inicial, número de aumentos de cupo y porcentaje de extracupo aumentan la probabilidad de que un cliente sea moroso de acuerdo a la clasificación obtenida en la tabla 31.

Por motivos de sigilo con la entidad, no se puede transcribir la ecuación del modelo más sin embargo recordarnos que la ecuación del modelo logístico se define como:

$$z = Logit\ (P) = ln\left(\frac{p}{1-p}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} x_1 + \boldsymbol{\beta_2} x_2 + \cdots + \boldsymbol{\beta_q} x_q$$

Tabla 32: Interpretación de Variables en base a los signos de los coeficientes de $\hat{\beta}$ del Modelo II de Entrenamiento

Variables	Descripción	Códigos	Signo	Explicación
Sexo	Mujer	1	Negativo	Clientes mujeres dismuyen la probabilidad de ser morosos
	Hombre	0		
Edad	Numérica	-	Negativo	La edad disminuye la probabilidad de ser morosos
Formalidad	Alta	1	Negativo	Clientes formales disminuyen la probabilidad de ser morsosos
	Otros	0		
Tiene Teléfono convencional	Posee Teléfono convencional	1	Negativo	Clientes formales disminuyen la probabilidad de ser morsosos
	No posee Teléfono convencional	0		
Tiene Teléfono celular	Posee Teléfono celular	1	Negativo	Clientes que poseen telefono celular dismuyen la probabilidad de ser morosos
	No posee teléfono celular	0		
Tiempo de Estabilidad domiciliaria	Numérica	-	Negativo	El tiempo de estabilidad domiciliaria disminuye la probabilidad de ser morosos
Provincia de Origen	CHI,ESM,RS,MO,PZA,PA,ZC,O,SO	1	Positivo	Las provincias codificadas descritas incrementan la probabilidad de ser morosos
	Otros	0		
Agencia de consumo	TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QO-ALM, QO-ALM5, QO-ALM8, STO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4	1	Positivo	Las agencias codificadas descritas incrementan la probabilidad de ser morosos
	Otros	0		
Estado actual de la tarjeta	Bloqueada	1	Positivo	Cliente con estado de Tc bloqueada incrementan la probabilidad de ser morosos
	Otros	0		
Motivo del bloqueo de la trajeta	Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración	1	Positivo	Clientes con motivo de bloqueo Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración incrementan la probabilidad de ser morosos
	Otros	0		
Cupo mensual disponible actual	Numérica	-	Negativo	El cupo mensual disponible actual disminuye la probabilidad de ser morosos
Cuota Mensual facturada	Numérica	-	Negativo	La cuota mensual facturada disminuye la probabilidad de ser morosos
Cupo mensual aprobado inicial	Numérica	-	Positivo	El cupo mensual aprobado inicial disminuye la probabilidad de ser morosos
No. aumentos de cupo	Numérica	-	Positivo	El número de aumentos de cupo disminuye la probabilidad de ser morosos
Porcentaje de Extracupo	Numérica	-	Positivo	El porcentaje de extracupo disminuye la probabilidad de ser morosos

Fuente: Elaboración Propia

Analizando el valor p, se puede evidenciar que las variables mayormente significativas son Sexo, estado de la tarjeta actual, motivo del bloqueo de la tarjeta actual, cupo mensual disponible actual, cuota mensual facturada, numero de aumentos de cupo, y porcentaje de extracupo dado que el valor p es menor a 0.001, sin embargo existen dos variables que son menos significativas como: edad y tiene teléfono convencional por lo tanto existe evidencia estadística para no rechazar la hipótesis nula de que los betas sean iguales a cero; mientras que el valor p para el resto de variables es inferior a 0.05 por lo tanto existe evidencia estadística para rechazar la hipótesis nula de que los $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ sean iguales a 0 lo cual indica que existe relación entre la variable dependiente y las covariables o variables explicativas.

En este caso no se descartarán las variables no significativas dado que el valor p es un valor referencial y no implica que al extraerlas el error de predicción disminuya por ende tomaremos en consideración las variables finalmente obtenidas como último paso del AIC.

3.2.1.4.1 TASAS DE CLASIFICACION DE PRECISION Y ERROR DEL MODELO II PARA REGRESIÓN LOGISTICA

Respecto a la matriz de confusión se procedió a clasificar a los clientes o sujetos de crédito utilizando como umbral óptimo $\boldsymbol{c}$ =0.45 con el objetivo de evitar el incremento de falsos positivos (Valor de z extremadamente <= 0.45) o negativos (Valor de z extremadamente >0.45) dentro de la clasificación, por ende, si el valor obtenido al reemplazar cada uno de los factores de la ecuación es menor que 0.45 se clasificará como no moroso, caso contrario se clasificara como moroso. De acuerdo a la tabla 32 se obtiene que, el 86.72% de los clientes no morosos fueron clasificados correctamente, mientras que el 13.28% fueron clasificados como morosos; y referente a los clientes morosos el 92.18% fueron clasificados correctamente y los restantes se encasillaron como no morosos.

Adicional se obtiene la precisión de predicción y de error que corresponden al 88.50% y 11.49% respectivamente en donde se justifica dicho nivel de error debido a las características de las variables que posee el modelo.

Tabla 33: Matriz de Confusión de Regresión Logística para Modelo II de Entrenamiento

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	1868	286	2154
Moroso (y=1)	82	966	1048
Total	1950	1252	3202

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	86,72%	13,28%	100,00%
Moroso (y=1)	7,82%	92,18%	100,00%

```
> precision3 <- sum(diag(confusion3))/sum(confusion3)
> precision3
[1] 0.8850718
> error.prediction3 <- 1-precision3
> error.prediction3
[1] 0.1149282
```

Fuente: Elaboración Propia

Finalmente se contrasta la matriz de confusión de entrenamiento y de prueba para evidenciar la precisión de la clasificación y error de predicción:

Tabla 34: Matriz de Confusión de Regresión Logística para Modelo II de Prueba

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	968	158	1126
Moroso (y=1)	52	546	598
Total	1020	704	1724

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	85,97%	14,03%	100,00%
Moroso (y=1)	8,70%	91,30%	100,00%

```
> precision3 <- sum(diag(confusion3))/sum(confusion3)
> precision3
[1] 0.8781903
> error.prediction3 <- 1-precision3
> error.prediction3
[1] 0.1218097
> #Grafica ROC
> library("pROC")
> g<-roc(Datos2$Default~prediccion3)
> g$auc
Area under the curve: 0.9337
```

Fuente: Elaboración Propia

3.2.2 ARBOLES DE DECISIÓN

3.2.2.1 ÁRBOLES DE DECISIÓN PARA MODELO I

Para evaluar el desempeño de clasificación de la muestra de desarrollo para el modelo I utilizamos la segunda metodología minería de datos "Árboles de decisión" para contrastar si en efecto esta técnica es la más adecuada para predecir la probabilidad de que un cliente sea moroso o no.
La figura 33, se ilustra el árbol final podado obtenido a través de la librería rpart de Rstudio. Esta librería utiliza el parámetro de complejidad (cp) con el objetivo de minimizar el error de cruce de validación, dado que mientras mayor sea el parámetro de complejidad, menores son las decisiones que contiene el árbol.

Con cp =0.01, el árbol de decisión final consta de cuatro nodos terminales. A continuación se detalla la descripción de cada uno de ellos (Ver figura 33):

- Nodo1: Muestra en resumen el total de clientes pertenecientes a la muestra de desarrollo, en donde la probabilidad de que un cliente sea bueno es del 67%, mientras que de ser malo 33%.

- Nodo2: Muestra a los clientes con formalidad tipo A (Clientes que se encuentra en relación de dependencia, es decir que sea asegurado o posea negocio propio con legalidad de RUC), los cuales representan el 59% de la submuestra total, e indica que el segmento descrito posea una probabilidad del 74% de ser buenos pagadores.

- Nodo3: Muestra a los clientes con otras formalidades (Clientes que realizan alguna actividad informal), los cuales representan el 41% de la submuestra total, e indica que aquellos sujetos que pertenezcan a este segmento tengan una probabilidad del 57% de ser buenos pagadores.

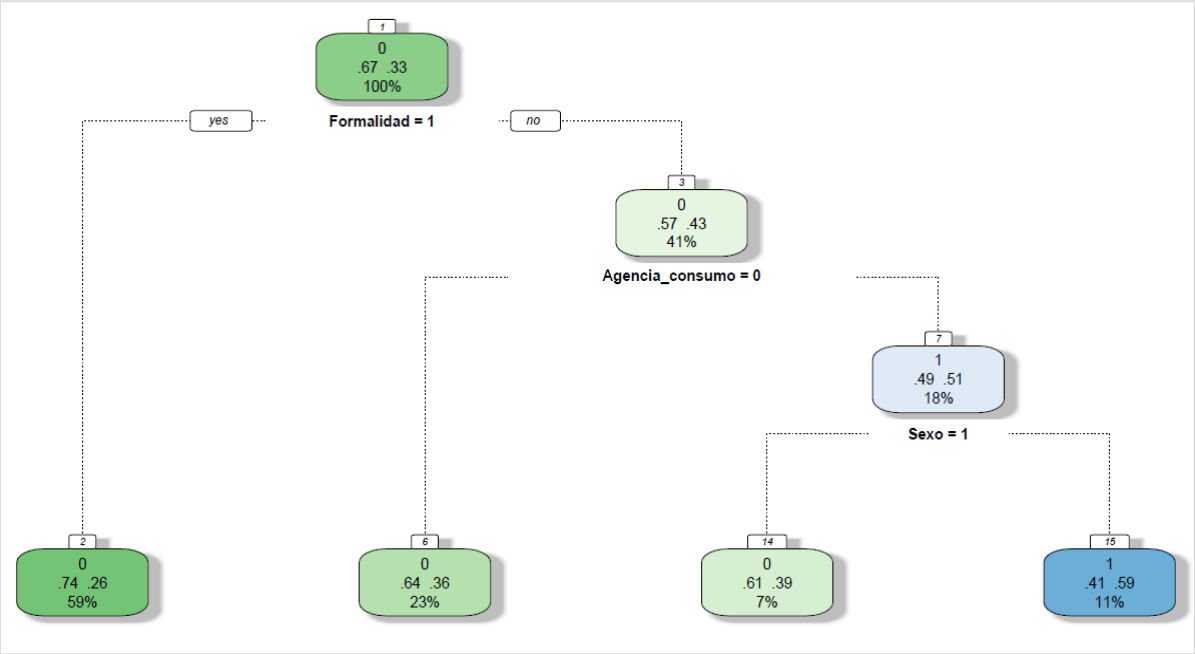
- Nodo6: Muestra en resumen los sujetos con otras formalidades pertenecientes a otras agencias diferentes de: TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QD-ALM, QO-ALM5, QO-ALM8, SO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4, e indica que los clientes que pertenezcan a este segmento tendrían una probabilidad del 64% de ser buenos pagadores.

- Nodo7: Muestra en resumen a los clientes con otras formalidades pertenecientes a las agencias de: TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QD-ALM, QO-ALM5, QO-ALM8, SO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4, e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser buenos pagadores del 51%.

- Nodo14: Muestra en resumen a clientes con otras formalidades pertenecientes a las agencias de: TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QD-ALM, QO-ALM5, QO-ALM8, SO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4 y de género femenino, que tengan una probabilidad de ser buenas pagadoras del 61%.

- Nodo15: Muestra en resumen a los sujetos con otras formalidades pertenecientes a las agencias de: TRN-ALM, QO-ALM2, GE-ALM5, GE-ALM2, QO-ALM1, DRA-ALM1, QD-ALM, QO-ALM5, QO-ALM8, SO-ALM, GE-ALM19, PO-ALM2, GE-ALM17, PO-ALM1, MTA-ALM2, QO-ALM4 y de género masculino, que tengan una probabilidad de ser buenos pagadores del 59%, es decir las mujeres son mejores pagadoras que los hombres.

Figura 21 : Árbol de decisión para Modelo I



Fuente: Elaboración Propia

3.2.2.1.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I PARA ÁRBOLES DE DECISIÓN

Respecto a la matriz de confusión de acuerdo a la figura 34 se obtiene que, el 93.18% de los clientes no morosos fueron clasificados correctamente, mientras que el 6.82% fueron clasificados como morosos; y referente a los clientes morosos el 20.04% fueron clasificados correctamente y los restantes se encasillaron como no morosos. Finalmente obteniendo las tasas de precisión de predicción y de error de 69.23% y 30.76% respectivamente, en donde se evidencia que el indicador de error supera al error obtenido en el modelo I de regresión logística.

Tabla 35: Matriz de Confusión de Árboles de Decisión para Modelo I de Entrenamiento

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	2007	147	2154
Moroso (y=1)	838	210	1048
Total	2845	357	3202

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	93,18%	6,82%	100,00%
Moroso (y=1)	79,96%	20,04%	100,00%

```
> precision2 <- sum(diag(confusion2))/sum(confusion2)
> precision2
[1] 0.6923798
> error.prediction2 <- 1-precision2
> error.prediction2
[1] 0.3076202
```

Fuente: Elaboración Propia

Finalmente se contrasta la matriz de confusión de entrenamiento y de prueba para evidenciar las tasas de precisión de la clasificación y error de predicción:

Tabla 36: Matriz de Confusión de Árboles de Decisión para Modelo I de Prueba

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	982	144	1126
Moroso (y=1)	376	222	598
Total	1358	366	1724

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	87,21%	12,79%	100,00%
Moroso (y=1)	62,88%	37,12%	100,00%

```
> precision <- sum(diag(confusion))/sum(confusion)
> precision
[1] 0.6983759
> error.prediction <- 1-precision
> error.prediction
[1] 0.3016241
```

Fuente: Elaboración Propia

3.2.2.2 ÁRBOLES DE DECISIÓN PARA MODELO II

Para el modelo II, utilizamos el mismo cp =0.01, para el cual realizaremos la explicación de los nodos finales prude obtenidos (Ver figura 36):

- Nodo1: Muestra en resumen el total de clientes pertenecientes a la muestra de desarrollo en donde la probabilidad de que un cliente sea bueno conforma el 67%, mientras que de ser malos pagadores el 33%.

- Nodo2: Muestra en resumen los clientes sin motivo de bloqueo, la cual representa el 58% de la muestra total, e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser buenos pagadores del 97%.

- Nodo3: Muestra en resumen a los sujetos con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser malos pagadores del 73%.

- Nodo6: Muestra en resumen a los sujetos con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración con cupo mensual asignado actual menor a $74 e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser malos pagadores del 63%.

- Nodo7: Muestra en resumen a los clientes con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración con cupo mensual asignado actual mayor e igual a $74 e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser malos pagadores del 91%.

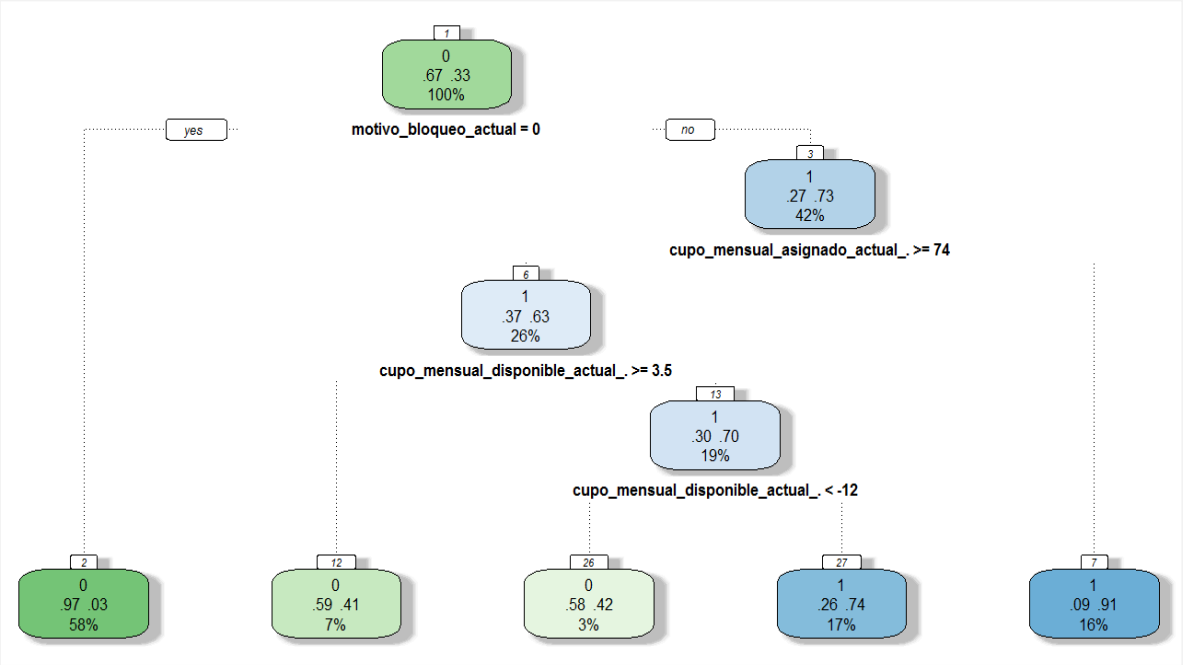
- Nodo12: Muestra en resumen a los sujetos con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración con cupo mensual asignado actual mayor e igual a $74 y con cupo disponible actual menor a $3,50 e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser buenos pagadores del 41%.

- Nodo13: Muestra en resumen a los sujetos con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración con cupo mensual asignado actual mayor e igual a $74 y con cupo disponible actual mayor e igual $3,50 e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser malos pagadores del 70%.

- Nodo26: Muestra en resumen a los sujetos con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración con cupo mensual asignado actual mayor e igual a $74 y con cupo disponible actual mayor e igual a $35 y menor a $-12 (extracupo, adicional de $12) e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser buenos pagadores del 58%.

- Nodo27: Muestra en resumen a los sujetos con motivo de bloqueo: Black Friday, Dación, Venta de cartera, Morosidad, Fallecimiento, Reestructuración con cupo mensual asignado actual mayor e igual a $74 y con cupo disponible actual menor a $-12 (extracupo) e indica que los clientes que pertenezcan a este segmento tengan una probabilidad de ser malos pagadores del 74%. Es decir a los clientes en mención no pueden asignárseles extracupo mayor a $12 dólares.

Figura 22: Árbol de decisión para Modelo II



Fuente: Elaboración Propia

3.2.2.2.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I PARA ÁRBOLES DE DECISIÓN

De acuerdo a la tabla 37 se obtiene que, el 91.27% de los clientes no morosos fueron clasificados correctamente, mientras que el 8.73% fueron clasificados como morosos; y referente a los clientes morosos el 83.21% fueron clasificados correctamente y los restantes se encasillaron como no morosos. Finalmente obteniendo las tasas de precisión de predicción y de error de 88.63% y 11.37% respectivamente se evidencia que el indicador de error es similar al indicador de error obtenido en el modelo II de regresión logística.

Tabla 37: Matriz de Confusión de Árboles de Decisión para Modelo II de Entrenamiento

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	1966	188	2154
Moroso (y=1)	176	872	1048
Total	2142	1060	3202

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	91,27%	8,73%	100,00%
Moroso (y=1)	16,79%	83,21%	100,00%

```
> precision <- sum(diag(confusion))/sum(confusion)
> precision
[1] 0.886321
> error.prediction <- 1-precision
> error.prediction
[1] 0.113679
```

Fuente: Elaboración Propia

Finalmente se contrasta la matriz de confusión de entrenamiento y de prueba para evidenciar la precisión de la clasificación y error de predicción:

Tabla 38: Matriz de Confusión de Árboles de Decisión para Modelo II de Prueba

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	1033	93	1126
Moroso (y=1)	112	486	598
Total	1145	579	1724

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	91,74%	8,26%	100,00%
Moroso (y=1)	18,73%	81,27%	100,00%

```
> precision2 <- sum(diag(confusion2))/sum(confusion2)
> precision2
[1] 0.8810905
> error.prediction2 <- 1-precision2
> error.prediction2
[1] 0.1189095
```

Fuente: Elaboración Propia

3.2.3 BAGGING

3.2.3.1 BAGGING PARA MODELO I

Para evaluar el desempeño de clasificación de la muestra de desarrollo para el modelo I utilizamos finalmente la cuarta metodología de "Bagging" o ensacado para contrastar si en efecto esta técnica es la más adecuada para predecir la probabilidad de que un cliente sea moroso o no.

Bagging, es una técnica de Machine Learning, el cual utiliza la técnica de remuestreo (bootstrapping) para mejorar la estabilidad o precisión de los algoritmos de aprendizaje de técnicas como clasificación estadística y regresión. Al igual que bosques aleatorios funciona como una caja negra y a diferencia de árboles de decisión evita el sobreajuste y reduce la varianza. Al utilizar la técnica de remuestreo, conlleva a que, en promedio, cada ajuste utilice aproximadamente 2/3 de las observaciones iniciales. Al tercio restante de la fracción se le denomina fuera de bolsa (OBB). Si por cada árbol ajustado de la metodología se registran las observaciones definidas, se puede predecir la respuesta de la observación $\boldsymbol{i}$,haciendo uso de aquellos árboles en los que la observación ha sido excluida (*OOB*) y promediándolos.

Continuando este proceso, se pueden obtener las predicciones para las k observaciones del *OOB-classification error*, dado que la variable de respuesta de cada observación se predice empleando únicamente los árboles en cuyo ajuste no conformó dicha observación, el *OOB-error* sirve como estimación del *test-error*.

De hecho, si el número de árboles es suficientemente alto, el *OOB-error* es prácticamente equivalente al *leave-one-out cross-validation error*. Esta es una ventaja añadida de los métodos de *bagging*, ya que evita tener que recurrir al proceso de *cross-validation* (computacionalmente costoso) para la optimización de los hiperparámetros por ende para el modelo I y II.

3.2.3.1.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I PARA BAGGING

De acuerdo a la figura 39 se tiene un OBB del 35.32% con una tasa de precisión del 64.68%. La tasa de error obtenida con respecto a las últimas metodologías es la más alta.

Figura 23: Tasa de Clasificación de error OBB del Modelo I de Desarrollo para Bagging

```
> print(mod)

Bagging classification trees with 25 bootstrap replicati
ons

Call: bagging.data.frame(formula = Default ~ ., data = D
atos, coob = TRUE)

Out-of-bag estimate of misclassification error:  0.3532
```

Fuente: Elaboración Propia

Posteriormente se contrasta los resultados de desarrollo vs validación, obteniendo tasas similares:

Figura 24: Tasa de Clasificación de error OBB del Modelo I de Validación para Bagging

```
> print(mod)

Bagging classification trees with 25 bootstrap replicati
ons

Call: bagging.data.frame(formula = Default ~ ., data = D
atos2, coob = TRUE)

Out-of-bag estimate of misclassification error:  0.3637
```

Fuente: Elaboración Propia

3.2.3.2 BAGGING PARA MODELO II

3.2.3.2.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I PARA BAGGING

Para el modelo II de la muestra de desarrollo, de acuerdo a la figura 41 se tiene un OBB del 14.02%, por ende su tasa de precisión es del 85.98%

Figura 25: Tasa de Clasificación de error OBB del Modelo II de Desarrollo para Bagging

```
> print(mod)

Bagging classification trees with 25 bootstrap replicati
ons

Call: bagging.data.frame(formula = Default ~ ., data = D
atos, coob = TRUE)

Out-of-bag estimate of misclassification error:  0.1402
```

Fuente: Elaboración Propia

Luego se contrastan los valores obtenidos de validación con desarrollo se evidencia que la tasa de error es similar, dado que el OBB del error es del 14.27%, obteniendo una tasa de precisión del 85.73%

Figura 26: Tasa de Clasificación de error OBB del Modelo II de Validación para Bagging

```
> print(mod)

Bagging classification trees with 25 bootstrap replicati
ons

Call: bagging.data.frame(formula = Default ~ ., data = D
atos2, coob = TRUE)

Out-of-bag estimate of misclassification error:  0.1427
```

Fuente: Elaboración Propia

3.2.4 BOSQUES ALEATORIOS

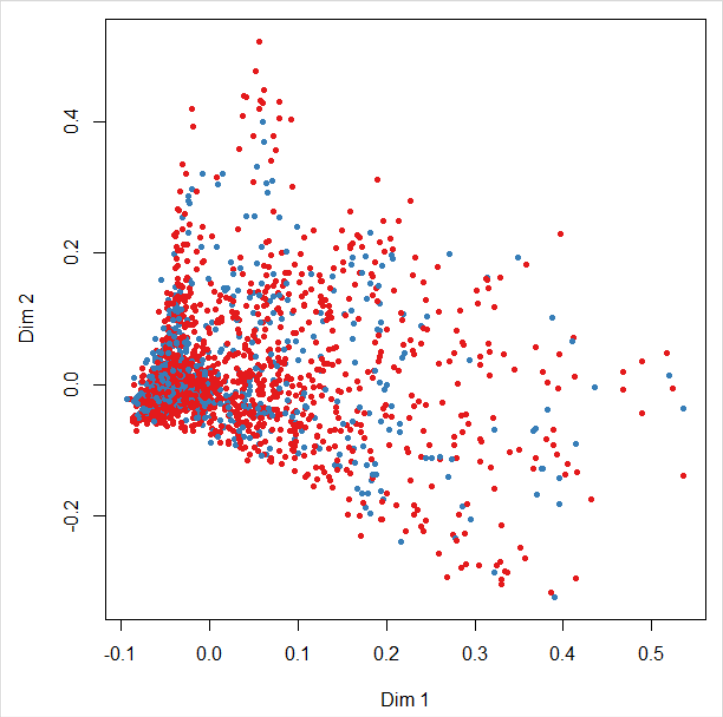
3.2.4.1 BOSQUES ALEATORIOS PARA MODELO I

Para evaluar el desempeño de clasificación de la muestra de desarrollo para el modelo I se emplea la tercera metodología de "Bosques aleatorios" para diferenciar si en efecto esta técnica es la más adecuada para predecir la probabilidad de que un cliente sea moroso o no.

Bosques aleatorios asigna de forma aleatoria una cantidad de variables explicativas para cada nodo. Una vez que se forman todos los árboles del bosque toma como referencia la media para efectuar la predicción al igual que árboles de decisión. A su vez funciona como una caja negra dado que al introducir la data se obtiene directamente la solución pero se desconocen los detalles de su funcionamiento por ende esta técnica impide obtener las reglas de predicción.

La importancia de la variable global se basa en la disminución de la media de precisión o Mean Decrease of Accuracy (MDA) sobre todas las predicciones cruzadas validadas fuera de la bolsa (OOB), cuando una variable dada se permuta después del entrenamiento, pero antes de la predicción la importancia de la variable global es la más popular, ya que es un número único por variable que se promedia sobre todas las predicciones.

Figura 27: Gráfica de Modelamiento Bosques Aleatorios

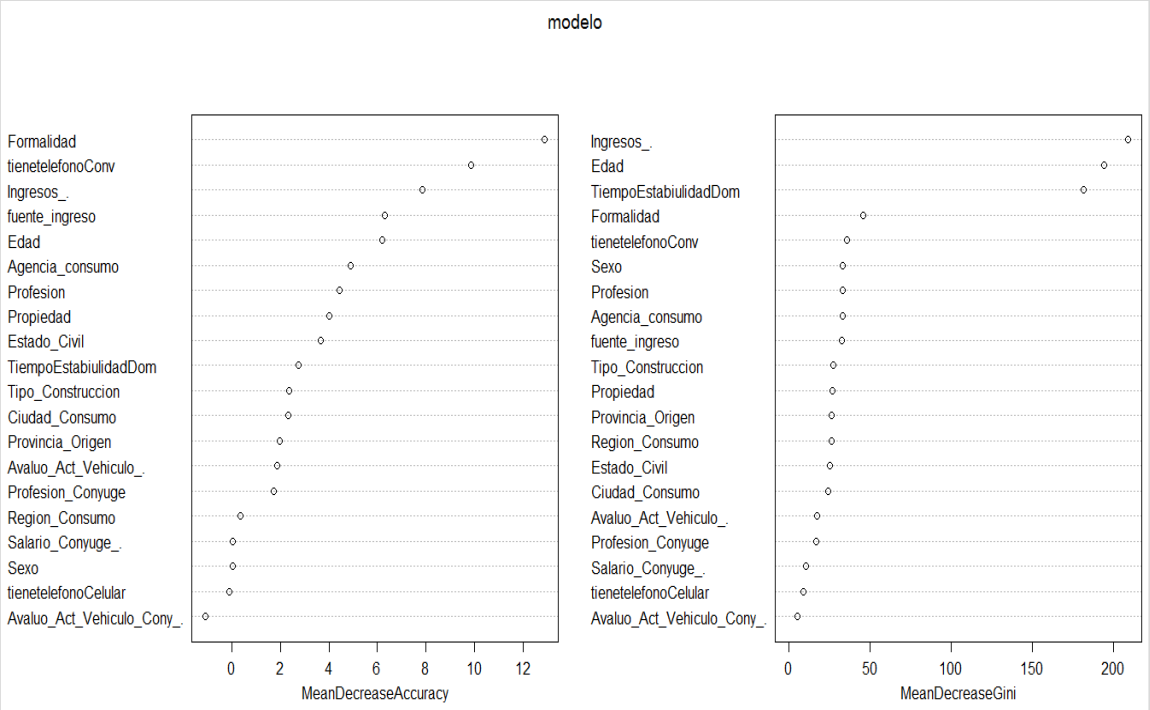


Fuente: Elaboración Propia

La importancia (Mean Drecrease GINI) mide la ganancia promedio de pureza por divisiones de una variable dada. Si la variable es útil, tiende a dividir nodos etiquetados mixtos en nodos de clase única puros. La división por variables permutadas no tiende a aumentar ni a disminuir las purezas del nodo. La importancia de GINI está relacionada con la función de decisión local que el bosque aleatorio usa para seleccionar la mejor división disponible y es relativamente más parcial, más inestable y tiende a responder a una pregunta más indirecta por ende tomaremos como referencia de clasificación la MDA.

La figura 44 contiene el ordenamiento de las variables por nivel importancia. En este caso las variables más representativas de acuerdo a la disminución de la media de precisión (Mean Decrease of Accuracy) son: Formalidad, si tiene teléfono convencional, Ingresos, fuente de ingresos y edad.

Figura 28: Variables Significativas del Modelo I mediante la Metodología de Bosques Aleatorios



Fuente: Elaboración Propia

3.2.4.1.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I PARA BOSQUES ALEATORIOS

De acuerdo a la figura 45 se obtiene que, el 87.93% de los clientes no morosos fueron clasificados correctamente, mientras que el 12.07% fueron clasificados como morosos; y referente a los clientes morosos el 26.34% fueron clasificados correctamente y los restantes se encasillaron como no morosos. Finalmente obteniendo las tasas de precisión de predicción y de error del 67.77% y 32.23% respectivamente se evidencia que el indicador de error es más alto en comparación al error obtenido en el modelo II de regresión logística y árboles de decisión.

Tabla 39: Matriz de Confusión de Bosques Aleatorios para Modelo I de Entrenamiento

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	1894	260	2154
Moroso (y=1)	772	276	1048
Total	2666	536	3202

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	87,93%	12,07%	100,00%
Moroso (y=1)	73,66%	26,34%	100,00%

```
> precision <- sum(diag(confusion))/sum(confusion)
> precision
[1] 0.67770
> error.prediction <- 1-precision
> error.prediction
[1] 0.32230
```

Fuente: Elaboración Propia

Por último, se comprueba la matriz de confusión de entrenamiento y de prueba para evidenciar las tasas de precisión de la clasificación y error de predicción:

Tabla 40: Matriz de Confusión de Bosques Aleatorios para Modelo I de Prueba

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	974	152	1126
Moroso (y=1)	387	211	598
Total	1361	363	1724

Observado	Predicción		
	No Moroso (y=0)	Moroso (y=1)	Total
No Moroso (y=0)	86,50%	13,50%	100,00%
Moroso (y=1)	64,72%	35,28%	100,00%

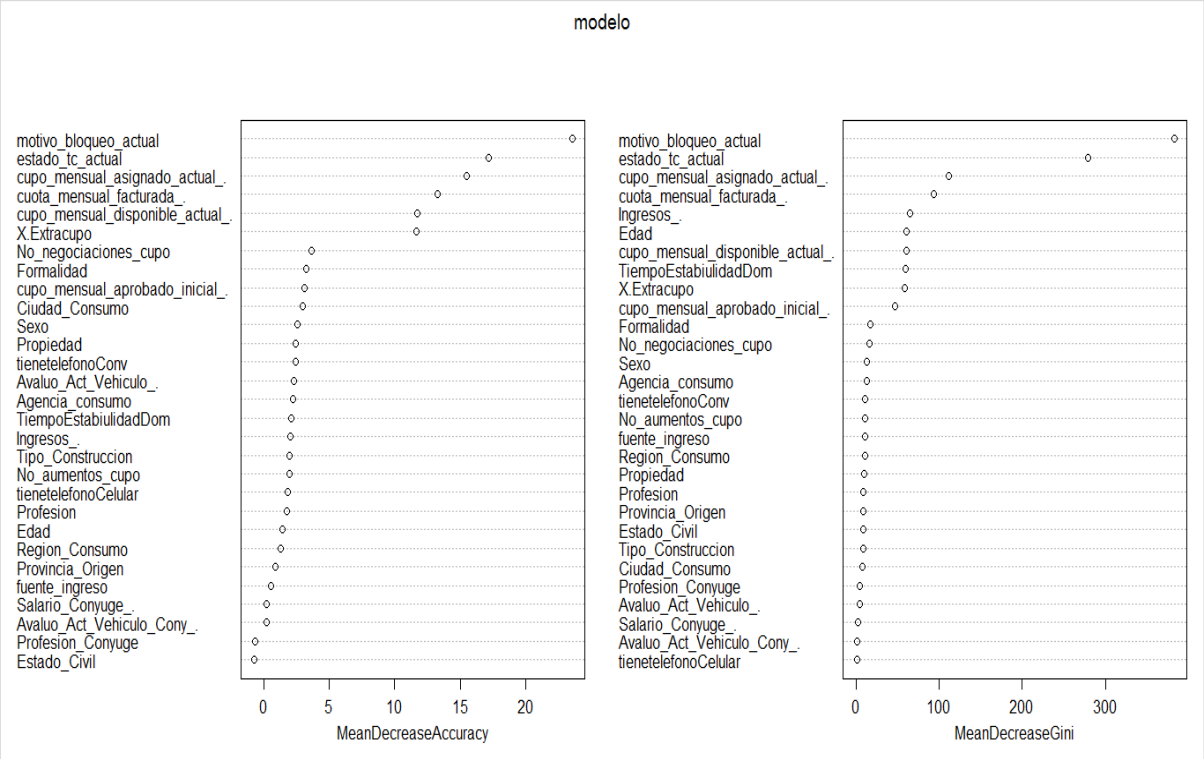
```
> precision <- sum(diag(confusion))/sum(confusion)
> precision
[1] 0.6873550
> error.prediction <- 1-precision
> error.prediction
[1] 0.3126450
```

Fuente: Elaboración Propia

3.2.4.2 BOSQUES ALEATORIOS PARA MODELO II

Para el modelo II de la muestra de desarrollo, de acuerdo a la figura 47 se evidencia el ordenamiento de las variables por nivel importancia. En este caso las variables más representativas de acuerdo a la disminución de la media de precisión (Mean Decrease of Accuracy) son: Motivo de bloqueo, estado de la tarjeta actual, cupo mensual asignado actual, cuota mensual facturada, cupo mensual disponible actual y porcentaje de extracupo.

Figura 29: Variables Significativas del Modelo II mediante la Metodología de Bosques Aleatorios



Fuente: Elaboración Propia

3.2.4.2.1 TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO II PARA BOSQUES ALEATORIOS

De acuerdo a la tabla 41 se obtiene que, el 89.14% de los clientes no morosos fueron clasificados correctamente, mientras que el 10.86% fueron clasificados como morosos; y referente a los clientes morosos el 85.78% fueron clasificados correctamente y los restantes se encasillaron como no morosos. Finalmente obteniendo las tasas de precisión de predicción y de error de 88.03% y 11.96% respectivamente se evidencia que el indicador de error es similar al indicador de error obtenido en el modelo II de regresión logística y árboles de decisión.

Tabla 41: Matriz de Confusión de Bosques Aleatorios para Modelo II de Entrenamiento

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	1920	234	**2154**
Moroso (y=1)	149	899	**1048**
Total	**2069**	**1133**	**3202**

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	**89,14%**	10,86%	**100,00%**
Moroso (y=1)	14,22%	**85,78%**	**100,00%**

```
> precision <- sum(diag(confusion))/sum(confusion)
> precision
[1] 0.8803873
> error.prediction <- 1-precision
> error.prediction
[1] 0.1196127
```

Fuente: Elaboración Propia

Se comprueba la matriz de confusión de desarrollo y de validación, para evidenciar las tasas de precisión de la clasificación y error de predicción:

Tabla 42: Matriz de Confusión de Bosques Aleatorios para Modelo II de Validación

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	1001	125	**1126**
Moroso (y=1)	94	504	**598**
Total	**1095**	**629**	**1724**

Observado	*Predicción*		
	No Moroso (y=0)	*Moroso (y=1)*	*Total*
No Moroso (y=0)	**88,90%**	11,10%	**100,00%**
Moroso (y=1)	15,72%	**84,28%**	**100,00%**

```
> precision <- sum(diag(confusion))/sum(confusion)
> precision
[1] 0.8729698
> error.prediction <- 1-precision
> error.prediction
[1] 0.1270302
```

Fuente: Elaboración Propia

3.3 RESUMEN DE MODELOS DATA MINING

3.3.1 COMPARATIVO DE LAS TASAS DE CLASIFICACIÓN DE PRECISION Y ERROR DEL MODELO I y II

En resumen de acuerdo a los errores de predicción y precisión obtenidos en la tabla 43 y 44, se evidencia que la metodología de regresión logística aplicada en el modelo I es la más adecuada para predecir el cumplimiento de pago de los sujetos a crédito dentro de la institución; mientras que para el modelo II (Variables demográficas + Comportamiento) se pueden utilizar las técnicas de regresión logística y árboles de decisión, pero dado que el error es aproximadamente cercano , se tomará en consideración regresión logística, debido a que dentro de esta metodología se pueden obtener los scores de crédito para la aprobación automática de los créditos.

Tabla 43: Resumen de Tasas de Error por Modelo

Desarrollo

Resultados	*Regresión Logística*	*Árboles de Decisión*	*Bagging*	*Bosques Aleatorios*
Modelo I	29,42%	30,76%	35,32%	32,23%
Modelo II	11,49%	11,37%	14,02%	11,96%

Validación

Resultados	*Regresión Logística*	*Árboles de Decisión*	*Bagging*	*Bosques Aleatorios*
Modelo I	30,68%	30,16%	36,37%	31,26%
Modelo II	12,18%	11,89%	14,27%	12,70%

Tabla 44: Resumen de Tasas de Precisión por Modelo

Desarrollo

Resultados	*Regresión Logística*	*Árboles de Decisión*	*Bagging*	*Bosques Aleatorios*
Modelo I	70,58%	69,24%	64,68%	67,77%
Modelo II	88,51%	88,63%	85,98%	88,04%

Validación

Resultados	*Regresión Logística*	*Árboles de Decisión*	*Bagging*	*Bosques Aleatorios*
Modelo I	69,32%	69,84%	63,63%	68,74%
Modelo II	87,82%	88,11%	85,73%	87,30%

Fuente: Elaboración Propia

CAPITULO 4

CONCLUSIONES Y RECOMENDACIONES

4.1. CONCLUSIONES

1. El comportamiento de pago de los clientes de la entidad es posible modelarlo a través de las metodologías matemáticas como son Regresión Logística, Arboles de decisión, Bagging y Bosques Aleatorios, en donde estos modelos se basan en el historial crediticio de 12 meses de desempeño. La robustez del modelo dependerá de que la información que lo integre sea completa y eficaz.

2. La empresa al no contar con punto de default definido para el segmento de estudio, se prevé que a partir de que el cliente entre mora mayor a 30 días existen altas probabilidades de que incumpla sus obligaciones crediticias con la entidad.

3. Las tasas de precisión obtenidos para las cuatros metodologías estadísticas: Regresión Logística, Arboles de Decisión, Bagging y Bosques Aleatorios, difieren para los modelos I y II, esto se debe a que el primer modelo al ser de otorgamiento, solo considera las variables que se ingresan en la solicitud de crédito (Variables demográficas), mientras que el modelo II, evalúa adicionalmente variables asociados al cumplimiento de pago y al ser más completo certifica el punto integridad, por ende posee mejor poder de discriminación y ordenamiento entre clientes buenos y malos.

4. En base al modelo I, la metodología de Regresión logística permitió encontrar el mejor modelo con menor tasa de error (29.42%) y mayor tasa de precisión en comparación a las metodologías de árboles de decisión(30.76%), Baagging (35.32%) y Bosques Aleatorios(32.23%), a pesar de ser estas dos últimas metodologías de aprendizaje automático o Machine Learning, no

resultaron ser eficientes para el pronóstico de este modelamiento, por lo cual se afirma que no existe una metodología única para determinar el mejor modelo, dado que depende de los tipo de variables que se ingresen al modelo. En este caso gran parte de las variables que contiene el modelo I son categóricas. El modelo I de regresión logística cuenta con el respaldo de los indicadores de calidad de K-S, GINI y ROC en base a las tasas de rendimiento obtenidas para modelos de originación.

5. Para el modelo II, la metodología de árboles de decisión permitió encontrar el mejor modelo con menor tasa de error (30.16%) y mayor tasa de precisión en comparación a las metodologías de regresión logística (30.68%), Bagging (36.37%) y Bosques Aleatorios (31.26%). En este caso, no salió resultante la metodología de regresión logística debido a que la diferenciación radica en tipo de variable que ingreso al modelo, es decir dicotómicas para regresión logística y continuas para aboles de decisión. En donde para arboles de decisión la librería rpart realiza el punto de corte basado en base al parámetro de complejidad (cp=0.01) para perfilar los clientes malos.

6. La elaboración del modelo de regresión logística depende de la experiencia y habilidades del analista de riesgo de crédito, en base al conocimiento de la obtención de la variable dependiente, puntos de corte, medidas de calidad y la toma de decisiones para la elección del modelo optimo, en cambio arboles de decisión, es una técnica más sencilla debido a que el software define internamente por default, el punto de corte.

7. Los modelos de otorgamiento son más penalizadores que los de comportamiento, al momento de puntuar a un cliente, debido a la falta de variables que aportan al poder predictivo del modelo, es decir un cliente puede ser puntuado en el modelo de otorgamiento con score 959, pero en el modelo II de comportamiento puede ser catalogado como 939.

8. El uso de las metodologías estadísticas aportan valor en base a estrategias diferenciadas permitiendo tomar decisiones eficaces con el objetivo de

fidelizar a los clientes y generar rentabilidad. En este caso cualquier modelo aportaría en comparación con el modelo genérico que utiliza la compañía dado que el error actual de predicción es del 50%.

4.2 RECOMENDACIONES

1. Se recomienda establecer reuniones de trabajo con los expertos del negocio del área de crédito y con el área tecnológica, con el objetivo de conocer el proceso de otorgamiento de crédito, las bases de datos, la codificación de variables internas, etc. para así identificar en conjunto las variables de mayor relevancia que aportaran al modelo.

2. En el caso de que los clientes se encuentren en mora mayor a 30 días de vencido se recomienda focalizar las estrategias en la gestión de la cobranza mediante visitas terrenas y gestión telefónica, con el objetivo de ofrecerles a los clientes facilidades de pago como refinanciaciones u otro tipo de negociación interna que posea la entidad para evitar deterioro en el tramo vencido.

3. Para el caso del modelo I y II, se recomienda que la ejecución del modelo I sea considerado siempre cuando el cliente obtenga crédito por primera vez con la entidad dado que inicialmente no contaremos con las variables de comportamiento, para ello el modelo II debe ser ejecutado una vez que el cliente posea experiencia mínima con la entidad de 3 meses a partir de la fecha de aprobación del crédito.

4. El modelo I, debe ser aplicado bajo la metodología de Regresión Logística, para ello se recomienda que su implementación interna se ejecute mediante interfaz web, con el objetivo de que el proceso de interoperabilidad sea optima al momento de la ejecución masiva de consultas.

5. Evaluando la implementación tecnológica para el modelo II, se recomienda utilizar el modelo de regresión logística dado que la tasa de error respecto a arboles de decisión es de apenas 1% y a su vez la implementación de la ecuación del modelo junto con los score agrega valor añadido en la automatización de consultas masivas en comparación a la visualización gráfica de Arboles de decisión.

6. Con el objetivo de evitar pérdidas en el otorgamiento y la gestión de cobranzas, se recomienda que para el Modelo I, los clientes que se puntúan con una calificación crediticia inferior a 781 sean rechazados dado que la probabilidad de que un cliente contraiga incumplimiento de pago supera el 50%, al igual que para el modelo II, si el cliente se puntúa por debajo de 219 se debe rechazar el ingreso de solicitudes dado que la probabilidades de incumplimiento de pago supera el 80%. Y a su vez para recuperar el segmento de clientes a rechazar, se recomienda realizar promociones y campañas exclusivas para aquellos clientes que se encuentren con score superior al punto de corte definido para cada modelo.

7. Se recomienda realizar pruebas de backtesting cada año a ambos modelos, dado que el punto de default que define a la variable dependiente puede cambiar en el tiempo. Es decir, actualmente el punto de default puede ser 30 días, en los próximos años puede ser 60 o 90, dependiendo de los factores internos (Políticas de Crédito) y externos (Factores Exógenos) que se susciten en la entidad.

BIBLIOGRAFÍA

- Abad, A. (2017-2018). Métodos de Regresión Lineal Simple, Múltiple y Logística. Guayaquil, Guayas, Ecuador.
- Abelraham, A. (2010). En *Applying Logistic Regression Model to The Second Primary Cancer Data* (págs. 105-210). Allahabad-India: Departament of Statistics, Mathematics, and Insurance, Faculty of Commerce, Ain Shams University, Egypt, Pushpa Publishing House.
- Abid, L., Masmoudi, A., & Zouari, S. (2016). The consumer loan's payment default predictive model: An application in a Tunisian comercial bank. *Asian Economic and Financial Review*, 27-42.
- Akaike, H. (1974). A new look at the statistical identification model. IEEE Control Systems Society.
- Breiman, F., Cutler, A., Liaw, R., & Wiener, M. (s.f.). *RDocumentation*. Obtenido de https://www.rdocumentation.org/packages/randomForest/versions/4.6-14
- Breiman, L. (1996). Machine Learning:Bagging Predictors. Springer US.
- Breiman, L. (2001). En M. L. Forest. Springer US.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Wadsworth, Belmonth, CA.
- Chen, S., Yeong, J., & Zone, D. S. (2014). A Hybrid Approach of Stepwise Regression, Logistic Regression, Support Vector Machine, and Decision Tree for Forecasting Fraudulent Financial Statements. *The Scientific World Journal*, 9.
- Everitt, B. (1998). *The Cambridge Dictionary of Statistics.* Cambridge University Press.
- Fisher, R. (1936). The use of multiple measurements in taxonomics problems, Annals of Eugenics. 179-188.
- Gaudencio, Z. (2008). *Probabilidad y Estadística Fundamentos y Aplicaciones.* Guayaquil, Ecuador: Primera Edición, Centro de Difusión y Publicaciones - Espol.
- Gutierrez, J., & Melo, L. (2011). *Pronostico de Incumplimiento de Pago.* Bogotá: Banco de la República.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. Data mining, Inference and Prediction.* Stanford, California: Springer Series in Statistics.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining , Inference and Prediction.* Springer.
- Hosmer, D., & Lemeshow, S. (s.f.). Applied Logistic Regression. New York: Second Edition,Wiley.
- IBM. (2007). *IBM Knowledge Center*. Obtenido de SPSS Statistics 22.0: https://www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/regression/logistic_regression_methods.htm
- Ihaka, R., & Gentleman, R. (1996). *R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics.*
- Iñiguez, C., & Bambino, C. (s.f.). *Selección de Perfiles de clientes mediante regresión logística para muestras desproporcionadas,validación,monitoreo y aplicación en la proyección de provisiones.* 2009.

- Li, J., Wei, H., & Hao, W. (2013). *Weight-Selected Attribute Bagging for Credit Scoring*. Beijing,China: Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology.
- Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest.
- Maindonald, J. H., & Braun, W. J. (2003-2017). *DAAG: Data Analysis and Graphics Data and Functions*. Obtenido de https://cran.r-project.org/web/randoms/DAAG/index.html
- Pino, P. (2017). Evaluación del Riesgo Crediticio mediante Árboles de Clasificación y Bosques Aleatorios. Sevilla.
- Rojo, H., & Miranda, M. (2009). *Cadenas de Markov.* Buenos Aires.
- Rubalcaba, J. (s.f.). *Modelización y Estadística en Biología*. Obtenido de Blog de WordPress.com: https://jgrubalcaba.wordpress.com/category/metodos-estadisticos/
- Spackman, K. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. San Mateo: Proc. Sixt Internat.Workshop on Machine Learning.
- Sweets, J., Dawes, R., & Monahan, J. (2000). Better decisions through science. Scientific American.
- Therneau, T., Atkinson, B., & Bryan,Rypley. (2017). *Recursive Partitioning and Regression Trees.* Obtenido de https://cran.r-project.org/web/packages/rpart/index.html
- Thomas, L. (2000). A survey of credit and behavioural scoring forecasting risk of lending to consumers. *International Journal of Forecasting*, (págs. 149-172). Edinburg.
- Thomas, L., Crook, J., & Edelman, D. (2002). *Credit Scoring and Its Applications.* Estados Unidos.
- Tomas, L. (2002). *Credit Scoring and Its Applications.* Estados Unidos.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory.Springer*. New York.
- Venables, B., Firth, D., Bates, D., Hornik, K., & Gebhardt, A. (1998). *RDocumentation*. Obtenido de https://www.rdocumentation.org/packages/MASS/versions/7.3-47
- Vera, M., Camanho, A., & Borgues, J. (2017). Predicting direct marketing response in banking: comparison of class imbalance methods. En *Service Business* (págs. 831-849). Springer Berlin Heidelberg.
- Vilar, J. (Junio de 2006). *Modelos Estadísticos Aplicados*. Obtenido de http://dm.udc.es/asignaturas/estadistica2/sec9_7.html